HIGHLY NONLINEAR MEASUREMENT ERROR MODELS IN NUTRITIONAL EPIDEMIOLOGY

A Dissertation

by

RUBIN WEI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Raymond J. Carroll
Committee Members,	Michael T. Longnecker
	Bani K. Mallick
	Nancy D. Turner
Head of Department,	Simon J. Sheather

May 2014

Major Subject: Statistics

ABSTRACT

This dissertation consists of two main projects in the area of measurement error models with application in nutritional epidemiology.

The first project studies the application of moment reconstruction and moment-adjusted imputation in the context of nonlinear Berkson-type measurement error. The idea of moment reconstruction and moment adjusted imputation, like regression calibration, is to replace the unobserved variable of interest which is subject to measurement error with a proxy, which can be used in a variety of subsequent analyses, without redoing the measurement error model each time a different downstream analysis is performed. However, both methods essentially require the homoscedastic classical measurement error model or non-classical model that can be easily reduced to a classical one. In the first project, we deal with a case where the measurement error structure is of nonlinear Berkson-type, and develop analogues of moment reconstruction and moment-adjusted imputation for this case. We use National Institutes of Health-AARP Diet and Health Study, where the latent variable is a dietary pattern score called the Healthy Eating Index-2005, and simulations to illustrate the methods. The numerical results show the promise of these methods in the nonlinear Berkson-type measurement error context.

In the second project, we consider measurement error models for two variables observed repeatedly and subject to measurement error. One variable is continuous but positive, while the other variable is a mixture of continuous and zero measurements. This second variable has two sources of zeros. The first source is episodic zeros, wherein some of the measurements for an individual may be zero and others positive. The second source is hard zeros, i.e., some individuals will always report

zero. An example is the consumption of alcohol from alcoholic beverages: some individuals consume alcoholic beverages episodically, while others never consume alcoholic beverages. However, with a small number of repeated measurements from individuals, it is not possible to determine those that are episodic zeros and those that are hard zeros. We develop a new measurement error model for this problem, and use Bayesian methods to fit it. We also contrast our approach for a single variable which is subject to excess zeros, with those methods that have been developed for a single variable and proven to be somewhat numerically unstable. Simulations and data analyses of two studies are used to show that the new method gives more realistic and numerically stable results than the maximum likelihood approach.

To My Family

ACKNOWLEDGEMENTS

First, I would like to thank my major advisor, Dr. Raymond J. Carroll, for his suggestions, guidance, patience, and continuous support over the years. It is indeed an honor to be Dr. Carroll's student. He has a busy schedule but he is always willing to help when I need his advice. He is such a wise man that I will always look up to.

I wish to thank Dr. Michael Longnecker, Dr. Bani Mallick, and Dr. Nancy Turner for serving on my committee and giving me timely feedback. In addition, I feel grateful for Dr. Longnecker's help at all phases of my study at Texas A&M and for keeping his door open to me and all the students all the time. I appreciate Dr. Cornelis J. Potgieter and Dr. Anindya Bhadra's collaboration with me on the projects. My graduate study at Texas A&M University is supported in part by a grant from the National Cancer Institute (R37-CA057030).

I want to thank Dr. Michael Speed for recruiting me to the department, offering me the technology teaching assistant and consulting project lead positions that support part of my study at Texas A&M, and continuous help, Dr. Edward Jones for his advice when I led distance learning students on consulting projects and when I participated in the 2012 Capital One Modeling Competition, and for his support, Ms. Kim Ritchie for showing me how to be a good team player.

I want to thank the faculty and the staff at the University of Tennessee, Knoxville (UTK) and Texas A&M University for making my studies and life enjoyable. I want to thank Drs. Frank Guess, Russell Zaretzki and Robert Mee for encouraging me to pursue a Ph.D. in Statistics. I am grateful that I met my friends from UTK and had a family atmosphere then and forever. I will always cherish the good times I had with my friends from Texas A&M.

Last but not least, I want to express my deepest gratitude to my parents for their love. They give me the best education and ever-lasting support. I want to thank my wife, for her deep love, continuous encouragement, and selfless support.

TABLE OF CONTENTS

LIST OF TABLES

TABLE Page

1. INTRODUCTION

We consider a measurement error problem. Let Y denote the outcome, $\mathbf{X}$ be the unobservable variables of interest, $\mathbf{Z}$ be there error-free covariates, and $\mathbf{W}$ be the observed but contaminated version of $\mathbf{X}$. It is assumed that

$$\mathbf{W} = \mathbf{X} + \mathbf{U}.$$

That is, the classical additive measurement error model, where $\mathbf{U}$ is independent of $(Y, \mathbf{X}, \mathbf{Z})$ and is thus homoscedastic.

Moment reconstruction (Freedman et al., 2004, 2008) aims to create an observable random variable, $\mathbf{X}_{\text{mr}}$, which has the same first two moments as the true $\mathbf{X}$ given Y, to substitute for the observable $\mathbf{W}$ in the downstream analyses. The aim of moment-adjusted imputation (Thomas et al., 2011) is to create a variable $\mathbf{X}_{\text{mai}}$ that has multiple moments that are the same as $\mathbf{X}$ and has the same covariance with $(Y, \mathbf{Z})$ as $\mathbf{X}$ has with $(Y, \mathbf{Z})$.

A major appeal of moment reconstruction and moment-adjusted imputation is that once the variable $\mathbf{X}_{\text{mr}}$ or $\mathbf{X}_{\text{mai}}$ is derived, it can be used in all downstream analyses. There is no need to redo a measurement error model each time a different downstream model is proposed. Indeed, Freedman et al. (2004) use the moment reconstructed variable, $\mathbf{X}_{\text{mai}}$, in logistic regression, linear discriminant analysis and in constructing a classification tree, simultaneously. Additionally, for example, if $\mathbf{X}$ is scalar and Y is binary, one might wish to model the effect of $\mathbf{X}$ on Y in a logistic regression with $\mathbf{X}$ modeled as linear, via a simple B-spline, or, following the typical epidemiological convention, as a step function, defined by either fixed pre-defined

categories or the quantiles of $\mathbf{X}$. Both moment reconstruction and moment-adjusted imputation are of course only approximate methods, but they have been shown to have good performance in a variety of areas.

The crucial constraint associated with moment reconstruction and moment adjusted imputation is that they are essentially restricted to the classical, homoscedastic measurement error model. This paper is concerned with the case that the error structure is that of a nonlinear, multivariate Berkson nature, so that, for example, for parameters $\boldsymbol{\Psi}$, individual-level random effects $\boldsymbol{\zeta} = \text{Normal}(0, \boldsymbol{\Sigma}_\zeta)$ with $\boldsymbol{\Sigma}_\zeta$ estimable, a known function $\mathcal{G}(\cdot)$, and for a sample with $i = 1, ..., n$,

$$\mathbf{X}_i = \mathcal{G}(\mathbf{W}_i, \mathbf{Z}_i, \boldsymbol{\Psi}, \boldsymbol{\zeta}_i). \tag{1.1}$$

The purpose of this research is to develop a method for model (1.1) that allows use of moment reconstruction and moment-adjusted imputation and that has good performance.

The research is motivated by the study of colorectal cancer Y in the National Institutes of Health-AARP Diet and Health Study (NIH-AARP) (Schatzkin et al., 2001; Reedy et al., 2008), with one of the risk predictors being the Healthy Eating Index-2005 (HEI-2005) (Guenther et al., 2008a,b), a multi-component index meant to measure adherence to the 2005 Dietary Guidelines for Americans. As described in Section 3.1, the HEI-2005 has a complex, heteroscedastic Berkson error structure of the type embodied by (1.1). Our aim is to derive methods in the same vein as moment reconstruction and moment-adjusted imputation, but in this very different context.

2. REVIEW OF MOMENT RECONSTRUCTION AND MOMENT-ADJUSTED IMPUTATION

Moment reconstruction (Freedman et al., 2004, 2008) aims to create an observable random variable, $\mathbf{X}_{\text{mr}}$, which has the same first two moments with true $\mathbf{X}$ given Y, to substitute for the observable $\mathbf{W}$ in the downstream analyses. The authors do not include the error-free covariates $\mathbf{Z}$, but $\mathbf{Z}$ is included in this report. It is assumed that $\mathbf{W}$ is an unbiased measurement of $\mathbf{X}$, that is,

$$E(\mathbf{W}|Y, \mathbf{Z}) = E(\mathbf{X}|Y, \mathbf{Z}).$$

Then the solution is given by

$$\mathbf{X}_{\text{mr}} = m(Y, \mathbf{Z})\{I - G(Y, \mathbf{Z})\} + \mathbf{W}G(Y, \mathbf{Z}).$$

where $m(Y, \mathbf{Z}) = E(\mathbf{W}|Y, \mathbf{Z})$ and $G(Y, \mathbf{Z}) = \{\text{cov}(\mathbf{W}|Y, \mathbf{Z})^{1/2}\}^{-1}\{\text{cov}(\mathbf{X}|Y, \mathbf{Z})\}^{1/2}$, $A^{1/2}$ is the symmetric square root of A.

The first and second conditional moments of $\mathbf{X}_{\text{mr}}$ and $\mathbf{X}$ are equal given $(Y, \mathbf{Z})$, i.e. $E(\mathbf{X}_{\text{mr}}|Y, \mathbf{Z}) = E(\mathbf{X}|Y, \mathbf{Z})$ and $\text{cov}(\mathbf{X}_{\text{mr}}|Y, \mathbf{Z}) = \text{cov}(\mathbf{X}|Y, \mathbf{Z})$. This implies that the unconditional second moments of $\mathbf{X}$ and $\mathbf{X}_{\text{mr}}$ are also equal.

The authors show that the moment reconstruction reduces to regression calibration in linear regression, without the normality assumption. It also yields a consistent estimator in logistic model with $\mathbf{X}|Y$ normally distributed, even with differential error conditional on Y. Moment reconstruction can be generalized to the scenario where $\mathbf{W}$ is biased for $\mathbf{X}$ but has a linear relationship, that is, if $E(\mathbf{W}|Y) = a(Y) + b(Y)E(\mathbf{X}|Y)$ with $a(Y)$ and $b(Y)$ known.

Moment-adjusted imputation (Thomas et al., 2011) aims to construct $\mathbf{X}_{\text{mai}}$, adjusted version of $\mathbf{W}$, so that $E(n^{-1}\sum_{i=1}^{n}\mathbf{X}_{\text{mai}}^{r}) = E(\mathbf{X}^{r}), r = 1,\ldots,M$. That is the first M sample moments of $\mathbf{X}_{\text{mai}}$ is an unbiased estimator of the first M moments of $\mathbf{X}$. The authors consider $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{X}_{\text{mai}}$ being univariate. Cross-moments between $\mathbf{X}$ and Y, $\mathbf{Z}$ can also be matched. Let $\mathcal{V} = (\mathbf{1}, \mathcal{Y}, \mathcal{Z})$ and its (i,k) element be V_{ik}, for $i = 1,\ldots,n$ and $k = 1,\ldots,K+2$, where $\mathbf{1}$ is a vector of ones, $\mathcal{Y} = (Y_1,\ldots,Y_n)^{\mathrm{T}}$, and $\mathcal{Z}$ be the n by K error-free covariates. In general, moment-adjusted imputation attempts to find $\mathbf{X}_{\text{mai}}$ with the property that $E(n^{-1}\sum_{i=1}^{n}\mathbf{X}_{\text{mai,i}}^{r}V_{ik}) = E(\mathbf{X}^{r}V_{k})$ for $r = 1,\ldots,M_k$. Matching moments of $\mathbf{X}$ is done by matching the first column of $\mathcal{V}$ and matching cross-moments between $\mathbf{X}$ and Y, $\mathbf{Z}$ is done by matching X with the rest columns of $\mathcal{V}$.

In simple linear regression, moment-adjusted imputation can replicate regression calibration and moment reconstruction by matching the first two moments and a cross-moment with the response, and the solution has a closed form. In logistic regression when Y is binary, moment-adjusted imputation can consistently estimate the parameters. The authors suggest to match four moments and two cross-moments. In logistic regression, moment-adjusted imputation does not require normality of the measurement error, and unlike moment reconstruction and regression calibration, still produces an unbiased estimator when the normality assumption is violated. For many nonlinear models, a closed form solution is not available. But the non-linearity can be well approximated by a lower-order polynomial. If the interest is in estimating the distribution of $\mathbf{X}$, a general recommendation (Thomas et al., 2011, page 1465) is to match the first four moments.

3. MOMENT RECONSTRUCTION AND MOMENT-ADJUSTED IMPUTATION IN NONLINEAR BERKSON MEASUREMENT ERROR MODELING

3.1 The NIH-AARP Study and the HEI-2005

3.1.1 The NIH-AARP Study

In this section, we describe how the model (1.1) can arise in practice.

For the National Institutes of Health-AARP Diet and Health Study (NIH-AARP), the outcome Y was incidence of colorectal cancer. We did separate analyses for men and women. Women with missing menopausal hormone therapy status were deleted because none of them developed colorectal cancer. In the main study, the sample sizes were $n = 293,615$ for men and $n = 198,245$ for women. There were $2,151$ men and 959 women who developed colorectal cancer. The covariates $\mathbf{Z}$ used were the same as in Reedy et al. (2008), consisting of age and dummy variable categories for education, ethnicity, body mass index, smoking status and physical activity, along with menopausal hormone therapy status for women. A food frequency questionnaire (FFQ) $\mathbf{Q}$ was obtained from all study participants.

The FFQ $\mathbf{Q}$ is known to be biased for usual nutritional intakes and also heteroscedastic, so that moment reconstruction and moment-adjusted imputation are not applicable for it. However, the NIH-AARP study has a small sub-study, known as a calibration study, in which 866 men and 854 women completed two 24-hour recalls. These recalls are assumed to be unbiased for usual dietary intake, although heteroscedastic. We will use this calibration study to model usual intakes, resulting in a form similar to (1.1).

3.1.2 HEI-2005

The Healthy Eating Index-2005 includes ratios of interrelated dietary components to energy and comprises 12 distinct component scores and a total summary score. Zhang et al. (2011b, Table 1) has a list of these components and the standards for scoring, and Guenther et al. (2008a,b) has further details. Intakes of each food or nutrient, represented by one of the 12 components, are expressed as a ratio to energy intake, evaluated, and assigned a score. The twelve HEI-2005 components represent 6 episodically consumed food groups (total fruit, whole fruit, total vegetables, dark green and orange vegetables and legumes or DOL, whole grains, and milk), 3 daily-consumed food groups (total grains, meat and beans, and oils), and 3 other daily-consumed dietary components (saturated fat, sodium, and calories from solid fats, alcoholic beverages and added sugars or SoFAAS). The important statistical aspect of the data is that out of the twelve food groups, six of them have excess zeros. Zhang et al. (2011b) report that among those children ages 2-8 in 2001-2004 National Health and Nutrition Examination Survey (NHANES), the percentages of 24HR-reported non-consumption of total fruit, whole fruit, whole grains, total vegetables, DOL and milk on any single day are 17%, 40%, 42%, 3%, 50% and 12%, respectively. The HEI-2005 is complex precisely because 6 of its twelve components are episodically consumed, thus making this a multivariate excess-zero problem.

The short-term dietary instruments used, the 24-hour recalls, are assumed to be unbiased measures of usual dietary intake on the original scale. However, they are not homoscedastic, so that the classical measurement error model does not hold for them. In any case, as described in Section 3.1.1, they are not available for the main NIH-AARP study. In addition, what is of interest is the HEI-2005 total score, which as seen in Zhang et al. (2011b, Table 1) is a highly nonlinear function of usual

intakes. Zhang et al. (2011b) uses the assumption of unbiasedness to model the usual intakes and hence to define and model the true HEI-2005. The important details of the model are given in Section 3.1.3, where we also justify (1.1). The modeling takes place in the calibration study with the 24-hour recalls.

3.1.3 The Model of Zhang et al. (2011b)

Using the 24-hour recall data, for each of the 6 episodically consumed food groups, two variables are defined: (a) whether a food from that group was consumed; and (b) the amount of the food that was reported on the 24-hour recall. For the 6 daily-consumed food groups and nutrients, only one variable indicating the consumption amount is defined. In addition, the amount of energy that is calculated from the 24-hour recall is of interest. The total number of dietary variables for each 24-hour recall is thus 12+6+1 = 19. The observed data are R_{ijk} for the i^{th} person, the j^{th} variable and the k^{th} replicate, $j = 1, \ldots, 19$ and $k = 1, 2$. Set $\mathbf{R}_{ik} = (R_{i1k}, ..., R_{i,19,k})^{\mathrm{T}}$, where

- $R_{i,2\ell-1,k}$ = Indicator of whether episodically consumed nutritional component # ℓ is consumed, with $\ell = 1, 2, 3, 4, 5, 6$.
- $R_{i,2\ell,k}$ = Amount of episodically consumed food # ℓ consumed. It equals to zero if food #ℓ is not consumed, with $\ell = 1, 2, 3, 4, 5, 6$.
- $R_{i,\ell+6,k}$ = Amount of daily consumed food or nutrient #ℓ, with $\ell = 7, 8, 9, 10, 11, 12$.
- $R_{i,19,k}$ = Amount of energy consumption as reported by the 24-hour recall.

Each of the 6 episodically consumed foods has 2 sets of latent variables, one for consumption and one for amount, while each of the 6 daily-consumed foods and nutrients as well as energy have 1 latent variable each, for a total of 19. The

latent random variables are $\boldsymbol{\zeta}_i = (\zeta_{i1}, ..., \zeta_{i,19}) = \text{Normal}(0, \boldsymbol{\Sigma}_\zeta)$, representing person-specific variation and $\boldsymbol{\xi}_{ik} = (\xi_{i1k}, ..., \xi_{i,19,k}) = \text{Normal}(0, \boldsymbol{\Sigma}_\xi)$, representing within-person variation. The $\boldsymbol{\zeta}_i$ and $\boldsymbol{\xi}_{ik}$ are mutually independent. As before, $\mathbf{Z}$ represents covariates while $\mathbf{Q}$ represents the food frequency questionnaire. In this model, food $\ell = 1, ..., 6$ being consumed on day k is equivalent to observing the binary $R_{i,2\ell-1,k}$, where

$$\begin{aligned} R_{i,2\ell-1,k} = 1 \quad &\Longleftrightarrow \quad S_{i,2\ell-1,k} \\ &= \quad (1, \mathbf{Q}_i^{\mathrm{T}}, \mathbf{Z}_i^{\mathrm{T}})\boldsymbol{\theta}_{2\ell-1} + \zeta_{i,2\ell-1} + \xi_{i,2\ell-1,k} > 0. \end{aligned} \tag{3.1}$$

Define the Box-Cox transformation as $g(y, \lambda) = (y^\lambda - 1)/\lambda$ for $\lambda \neq 0$ and $= \log(y)$ if $\lambda = 0$. If the food is consumed, we model the amount reported, $R_{i,2\ell,k}$, as

$$\begin{aligned} [g_{\text{tr}}(R_{i,2\ell,k}, \lambda_\ell) | R_{i,2\ell-1,k} = 1] \quad &= \quad S_{i,2\ell,k} \\ &= \quad (1, \mathbf{Q}_i^{\mathrm{T}}, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{\theta}_{2\ell} + \zeta_{i,2\ell} + \xi_{i,2\ell,k}, \end{aligned} \tag{3.2}$$

where $g_{\text{tr}}(y, \lambda) = \sqrt{2}\{g(y, \lambda) - \mu(\lambda)\}/\sigma(\lambda)$, and $\{\mu(\lambda), \sigma(\lambda)\}$ are the sample mean and standard deviation of $g(y, \lambda)$, computed from the nonzero food data. This standardization is a convenient device to improve the numerical performance of the algorithm without affecting conclusions.

The reported consumption of daily consumed foods or nutrients, plus energy, $\ell = 7, \ldots, 13$ is modeled as

$$g_{\text{tr}}(R_{i,\ell+6,k}, \lambda_\ell) \quad = \quad S_{i,\ell+6,k} = (1, \mathbf{Q}_i^{\mathrm{T}}, \mathbf{Z}_i^{\mathrm{T}})\boldsymbol{\theta}_{\ell+6} + \zeta_{i,\ell+6} + \xi_{i,\ell+6,k}, \tag{3.3}$$

where $g_{\rm tr}(y,\lambda) = \sqrt{2}\{g(y,\lambda) - \mu(\lambda)\}/\sigma(\lambda)$, and $\{\mu(\lambda), \sigma(\lambda)\}$ are the sample mean and standard deviation of $g(y,\lambda)$, computed from the data. As seen in (3.2)-(3.3), different transformations $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_{13})^{\rm T}$ are used for the different types of dietary components.

Denote the collection of $\boldsymbol{\theta}_j$ as $\boldsymbol{\Theta}$. Zhang et al. (2011b) use MCMC to estimate $(\boldsymbol{\Theta}, \boldsymbol{\Sigma}_\zeta, \boldsymbol{\Sigma}_\xi)$. From that, usual intake and the usual HEI-2005 component scores are defined as follows. Consider the first episodically consumed dietary component, a food group. Since the 24-hour recalls are unbiased for a person's usual intake, the usual intake is the expectation of the reported intake conditional on the person's random effects $\boldsymbol{\zeta}_i$. Let $g_{\rm tr}^{-1}(\cdot)$ be the inverse transformation of $g_{\rm tr}(\cdot)$, and let $\Phi(\cdot)$ be the standard normal distribution function. Then, a person's usual intake of the first episodically consumed dietary component is

$$\begin{aligned} X_{i1,\rm com} &= X_{i1,\rm com}(\mathbf{Q}_i, \mathbf{Z}_i, \boldsymbol{\Theta}, \boldsymbol{\Sigma}_\zeta, \boldsymbol{\Sigma}_\xi, \boldsymbol{\zeta}_i) \\ &= E(R_{i2}|\mathbf{Q}_i, \mathbf{Z}_i, \boldsymbol{\Theta}, \boldsymbol{\Sigma}_\zeta, \boldsymbol{\Sigma}_\xi, \zeta_{i1}, \zeta_{i2}) \\ &= \Phi\{(1, \mathbf{Q}_i^{\rm T}, \mathbf{Z}_i^{\rm T})\boldsymbol{\theta}_1 + \zeta_{i1}\} E\left[g_{\rm tr}^{-1}\left\{(1, \mathbf{Q}_i^{\rm T}, \mathbf{Z}_i^{\rm T})\boldsymbol{\theta}_2 + \zeta_{i2} + \xi_{i21}, \lambda_1\right\} |\boldsymbol{\zeta}_i\right]. \end{aligned}$$

Some remedies are used to make the expectation computable, but the details are not of interest here. Usual intake for the other episodically consumed food groups is defined in the same manner, and similarly for the daily consumed components, which do not have the leading term involving the standard normal distribution function. The collection of terms $(X_{ij,\rm com})_{j=1}^{13}$ is denoted as $\mathbf{X}_{i,\rm com}$.

The end result of this process is that the true HEI-2005 total score, X_T , and true energy, X_E have the representations that for functions $\mathcal{G}_T$ and $\mathcal{G}_E$,

$$X_T = \mathcal{G}_T(\mathbf{Q}, \mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}_\zeta, \boldsymbol{\Sigma}_\xi, \boldsymbol{\zeta}); \tag{3.4}$$

$$X_E = \mathcal{G}_E(\mathbf{Q}, \mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}_\zeta, \boldsymbol{\Sigma}_\xi, \boldsymbol{\zeta}), \tag{3.5}$$

where $\boldsymbol{\zeta} = \text{Normal}(0, \boldsymbol{\Sigma}_\zeta)$ is independent of $(\mathbf{Q}, \mathbf{Z})$. Setting $\mathbf{X} = (X_T, X_E)^{\mathrm{T}}$, we write

$$\mathbf{X} = \mathcal{G}(\mathbf{Q}, \mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}_\zeta, \boldsymbol{\Sigma}_\xi, \boldsymbol{\zeta}) = \{\mathcal{G}_T(\mathbf{Q}, \mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}_\zeta, \boldsymbol{\Sigma}_\xi, \boldsymbol{\zeta}), \mathcal{G}_E(\mathbf{Q}, \mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}_\zeta, \boldsymbol{\Sigma}_\xi, \boldsymbol{\zeta})\}^{\mathrm{T}}, \tag{3.6}$$

which is the specific form of (1.1) for this application.

3.2 Methods

3.2.1 Basic Approach

Our basic approach to constructing analogues of moment reconstruction and moment-adjusted imputation is to transform the data into a form amenable for these methods. Recall (3.6) and define $\mathbf{W}$ as the vector of FFQ measurements for the total score and energy: it is of course a function of $\mathbf{Q}$. We require that $\mathbf{X}$ and $\mathbf{W}$ have the same number of components p_x; in our case, $p_x = 2$. For a vector of parameters $\boldsymbol{\lambda}$ of length p_x, let $g(\mathbf{X}, \boldsymbol{\lambda})$ be the componentwise Box-Cox transformations for $1-$ Total Score/100 and Energy/2500, the former for the left skewness of the total score and the division for the convenience when estimating parameters. We first assume that there are parameters $(\boldsymbol{\lambda}_w, \boldsymbol{\lambda}_x)$ with the property that

$$g\left(\mathbf{W}, \boldsymbol{\lambda}_w\right) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1^{\mathrm{T}} g\left(\mathbf{X}, \boldsymbol{\lambda}_x\right) + \boldsymbol{\beta}_2^{\mathrm{T}} \mathbf{Z} + \mathbf{U}, \tag{3.7}$$

where $\boldsymbol{\beta}_1$ is of full rank and $\mathbf{U} = \text{Normal}\,(0, \boldsymbol{\Sigma}_u)$ is independent of $(Y, \mathbf{X}, \mathbf{Z})$.

Next, to relate $\mathbf{X}$ to $\mathbf{Z}$, we assume that

$$g(\mathbf{X}, \boldsymbol{\lambda}_x) = \boldsymbol{\alpha}_0^{\mathrm{T}} + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{Z} + \mathbf{V}, \tag{3.8}$$

where $\mathbf{V} = \text{Normal}(0, \boldsymbol{\Sigma}_v)$ is independent of $(\mathbf{Z}, \mathbf{U})$. Define $\widetilde{\mathbf{X}} = g(\mathbf{X}, \boldsymbol{\lambda}_x)$, $\widetilde{\mathbf{U}} = \left(\boldsymbol{\beta}_1^{\mathrm{T}}\right)^{-1}\mathbf{U}$ and $\widetilde{\mathbf{W}} = \left(\boldsymbol{\beta}_1^{\mathrm{T}}\right)^{-1}\left\{g\left(\mathbf{W}, \boldsymbol{\lambda}_w\right) - \boldsymbol{\beta}_0 - \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{Z}\right\}$. We can rewrite (3.7) as

$$\widetilde{\mathbf{W}} = \widetilde{\mathbf{X}} + \widetilde{\mathbf{U}}, \tag{3.9}$$

where $\widetilde{\mathbf{U}}$ is independent of $\mathbf{Z}$ and has covariance matrix $\boldsymbol{\Sigma}_{\widetilde{u}}$. With this construction, we now have a scenario where moment reconstruction and moment-adjusted imputation can be applied directly, since (3.9) is a classical measurement error model.

3.2.2 Estimating the Parameters in Section 3.2.1

Remember that $\mathbf{W}$ is a function of $\mathbf{Q}$. Recall equations (3.4)-(3.6) and (3.7)-(3.9). Estimation of the transformation parameters $(\boldsymbol{\lambda}_w, \boldsymbol{\lambda}_x)$ is required. In general, this would not be possible since we do not observe $(\mathbf{Z}, \mathbf{Q}, \mathbf{W}, \mathbf{X})$ even on a subset of the data. However, since $\boldsymbol{\zeta}$ is independent of $(\mathbf{Z}, \mathbf{Q}, \mathbf{W})$, by generating realizations of $\boldsymbol{\zeta}$ and substituting into (3.6), we can generate $(\mathbf{Z}, \mathbf{Q}, \mathbf{W}, \mathbf{X}^\star)$ that have the same joint distribution as $(\mathbf{Z}, \mathbf{Q}, \mathbf{W}, \mathbf{X})$. Parameter estimates can therefore be obtained easily from these simulated random variables.

We outline here how the transformation parameters were estimated using said pairs. Let $\widehat{\boldsymbol{\alpha}}_0\left(\boldsymbol{\lambda}_x\right)$ and $\widehat{\boldsymbol{\alpha}}_1\left(\boldsymbol{\lambda}_x\right)$ denote the least squares parameter estimates when performing a linear regression of $g(\mathbf{X}^\star, \boldsymbol{\lambda}_x)$ on $\mathbf{Z}$ for a fixed value of $\boldsymbol{\lambda}_x$. Define residuals $\mathbf{V}^\star\left(\boldsymbol{\lambda}_x\right) = g(\mathbf{X}^\star, \boldsymbol{\lambda}_x) - \widehat{\boldsymbol{\alpha}}_0\left(\boldsymbol{\lambda}_x\right) - \widehat{\boldsymbol{\alpha}}_1\left(\boldsymbol{\lambda}_x\right)\mathbf{Z}$. Since the distribution of $\mathbf{V}^\star$ is assumed Gaussian for the true value of $\boldsymbol{\lambda}_x$, the estimated transformation parameter

$\widehat{\boldsymbol{\lambda}}_x$ is, component-wise, the value that maximizes the absolute correlation between the percentiles of $\mathbf{V}^\star$ and the percentiles of the standard Gaussian distribution. A similar procedure is used to estimate $\boldsymbol{\lambda}_w$, and then the other parameters.

3.2.3 Moment-Adjusted Imputation

Model (3.9) is exactly a classical measurement error model, to which moment-adjusted imputation can be applied. In principle, one has to do a bivariate moment-adjusted imputation, for which programs are not yet available. However, in our context, the HEI-2005 total score is very nearly independent of energy intake, and thus for simplicity we used the programs mentioned in Thomas et al. (2011) separately for HEI-2005 total score and energy.

3.2.4 Moment Reconstruction

Define $m(Y, \mathbf{Z}) = E(\widetilde{\mathbf{W}}|Y, \mathbf{Z})$ and $G(Y, \mathbf{Z}) = \{\text{cov}(\widetilde{\mathbf{W}}|Y, \mathbf{Z})^{1/2}\}^{-1}\{\text{cov}(\widetilde{\mathbf{X}}|Y, \mathbf{Z})\}^{1/2}$. Moment reconstruction now proceeds by substituting the unobserved $\widetilde{\mathbf{X}}$ by $\widetilde{\mathbf{X}}_{\text{mr}} = m(Y, \mathbf{Z})\{I - G(Y, \mathbf{Z})\} + \widetilde{\mathbf{W}}G(Y, \mathbf{Z})$ which has been constructed so that the first two conditional moments of $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{X}}_{\text{mr}}$ are equal. Of course, to get to this point, the additional parameters $(\boldsymbol{\lambda}_w, \boldsymbol{\lambda}_x)$ and $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$ also need to be estimated, see Section 3.2.1. The model of interest is assumed to be a function of $\widetilde{\mathbf{X}}$, with the function known up to a vector of parameters.

In any given example of moment reconstruction, constructing $m(Y, \mathbf{Z})$ and $G(Y, \mathbf{Z})$ is done on a case-by-case basis. Freedman, et al. (2004) show how to do this explicitly if there are no additional covariates $\mathbf{Z}$, and if Y is binary as in logistic regression. Specifically, $m(Y, \mathbf{Z})$ is the mean of $\widetilde{\mathbf{W}}$ among those sharing the same values of Y, and $\text{cov}(\widetilde{\mathbf{X}}|Y, \mathbf{Z})$ is the covariance of $\widetilde{\mathbf{W}}$ among those sharing the same values of Y minus $\text{cov}(\widetilde{\mathbf{U}})$. In the example of Section 3.3, however, $\mathbf{Z}$ is of dimension > 20 and thus this simple device is not applicable. Instead, we used

the following device. Using the parameters estimates found in Section 3.2.2, define $\widetilde{\mathbf{X}}^{\star}_{\dagger} = g(\mathbf{X}^{\star}, \widehat{\boldsymbol{\lambda}}_x)$ and $\widetilde{\mathbf{W}}_{\dagger} = (\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}})^{-1}\{g(\mathbf{W}, \widehat{\boldsymbol{\lambda}}_w) - \widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}_2^{\mathrm{T}}\mathbf{Z}\}$. The estimate $\widehat{m}(Y, \mathbf{Z})$ of $m(Y, \mathbf{Z}) = E(\widetilde{\mathbf{W}}|Y, \mathbf{Z})$ is found by performing separate linear regressions of $\widetilde{\mathbf{W}}_{\dagger}$ on the covariates $\mathbf{Z}$ for both the cases ($Y = 1$) and controls ($Y = 0$). In estimating the covariance component, we assume that $\mathrm{cov}(\widetilde{\mathbf{W}}|Y, \mathbf{Z}) = \mathrm{cov}(\widetilde{\mathbf{X}}|Y, \mathbf{Z}) + \mathrm{cov}(\widetilde{\mathbf{U}})$. We also assume that $\mathrm{cov}(\widetilde{\mathbf{W}}|Y, \mathbf{Z})$ only depends on $\mathbf{Z}$ through $m(Y, \mathbf{Z})$. The estimate $\widehat{\mathrm{cov}}(\widetilde{\mathbf{W}}|Y, \mathbf{Z})$ is found by calculating the residuals $\widetilde{\mathbf{W}}_{\dagger} - \widehat{m}(Y, \mathbf{Z})$ in both the cases and controls, and then finding the covariance matrices corresponding to those residuals. While we are unable to estimate $\mathrm{cov}(\widetilde{\mathbf{X}}|Y, \mathbf{Z})$ directly from the data, we are able to find estimates of both $\mathrm{cov}(\widetilde{\mathbf{W}}|Y, \mathbf{Z})$ and $\mathrm{cov}(\widetilde{\mathbf{U}})$. Define residuals $\widetilde{\mathbf{U}}_{\dagger i} = \widetilde{\mathbf{W}}_{\dagger i} - \widetilde{\mathbf{X}}^{\star}_{\dagger i}$ and let $\widehat{\boldsymbol{\Sigma}}_{\widetilde{u}}$ be the sample covariance matrix of the $\widetilde{\mathbf{U}}_{\dagger i}$, the estimate of $\mathrm{cov}(\widetilde{\mathbf{U}})$. Then $\widehat{\mathrm{cov}}(\widetilde{\mathbf{X}}|Y, \mathbf{Z}) = \widehat{\mathrm{cov}}(\widetilde{\mathbf{W}}|Y, \mathbf{Z}) - \widehat{\boldsymbol{\Sigma}}_{\widetilde{u}}$.

3.2.5 Regression Calibration

Regression calibration is defined as replacing a latent variable by its expectation given the observed covariates. We do this in the original data scale, as follows. We use the characterization $\mathbf{X} = \mathcal{G}(\mathbf{Q}, \mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}_{\zeta}, \boldsymbol{\Sigma}_{\xi}, \boldsymbol{\zeta})$ given in (3.6). We compute $E(\mathbf{X}|\mathbf{Q}, \mathbf{Z})$ by Monte-Carlo. Set $B = 500$, and generate $(\boldsymbol{\zeta}_{1,\mathrm{rc}}, ..., \boldsymbol{\zeta}_{B,\mathrm{rc}}) = \mathrm{Normal}(0, \boldsymbol{\Sigma}_{\zeta})$. Let the estimates of $(\boldsymbol{\Theta}, \boldsymbol{\Sigma}_{\zeta}, \boldsymbol{\Sigma}_{\xi})$ be $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Sigma}}_{\zeta}, \widehat{\boldsymbol{\Sigma}}_{\xi})$. Then $\widehat{E}(\mathbf{X}|\mathbf{Q}, \mathbf{Z}) = \widehat{\mathbf{X}}_{\mathrm{rc}} = B^{-1}\sum_{b=1}^{B}\mathcal{G}(\mathbf{Q}, \mathbf{Z}, \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Sigma}}_{\zeta}, \widehat{\boldsymbol{\Sigma}}_{\xi}, \boldsymbol{\zeta}_{b,\mathrm{rc}})$.

3.3 The NIH-AARP Study Analysis

3.3.1 Overview

The data are described in Section 3.1.1. We fit the data using logistic regression: results were very similar for Cox regression, where, following Thomas et al. (2011), $\mathbf{Z}$ was augmented by case-control status when constructing moment reconstruction and moment-adjusted imputation.

- We did 5 different analyses with 3 different models. The analyses were (a) use of the FFQ in the original scale and ignoring measurement error; (b) regression calibration (RC) on the original scale as described in Section 3.2.5; (c) moment reconstruction (MR); (d) moment-adjusted imputation (MAI); and (e) Monte-Carlo maximum likelihood (MCML) on the original scale, with the score functions computed using $B = 500$ simulations.

- The 3 different models were (i) linear logistic regression; (ii) quadratic B-spline with 4 basis functions because there was some hint of curvature in the regression model when using the FFQ; and (iii) dummy variable regression for the HEI-2005 total score based on the estimated quintiles of the true total score. For (iii), because all transformations are monotone, the quintiles in the transformed scale are immediate. For men, the quintile break points are $(50.6, 58.0, 64.0, 70.3)$, while for women they are $(55.9, 62.9, 68.3, 73.7)$.

- When evaluating (i) and (ii), we computed the relative risk when moving from a true total score of 45, representing a poor diet, to a true total score of 75, representing a very good diet. When evaluating (iii), we computed relative risk between the first and fifth quintile, also representing a change from a poor diet to a good diet.

The computation is implemented in MATLAB.

3.3.2 Results

Results for the analysis of the HEI-2005 Total Score are provided in Table 3.1.

Consider first the analysis for men. With one exception, discussed below, the relative risks are consistent within method. For the linear risk model and the spline model, moment-adjusted imputation, regression calibration and Monte-Carlo max-

		Men				Women			
		RR	p-value	L 95%	H 95%	RR	p-value	L 95%	H 95%
MR									
	Lin	**0.699**	0.000	0.614	0.796	**0.767**	0.005	0.637	0.925
	Quin	**0.710**	0.000	0.613	0.822	**0.790**	0.029	0.639	0.976
	Spl	**0.725**	0.000	0.634	0.830	**0.729**	0.002	0.600	0.887
MAI									
	Lin	**0.652**	0.000	0.577	0.736	**0.712**	0.000	0.595	0.852
	Quin	**0.656**	0.000	0.570	0.754	**0.749**	0.006	0.609	0.922
	Spl	**0.663**	0.000	0.583	0.755	**0.706**	0.000	0.583	0.853
RC									
	Lin	**0.651**	0.000	0.555	0.764	**0.647**	0.004	0.481	0.870
	Quin								
	Spl	**0.661**	0.000	0.561	0.779	**0.676**	0.024	0.481	0.950
FFQ									
	Lin	**0.731**	0.000	0.650	0.822	**0.832**	0.053	0.691	1.002
	Quin	**0.723**	0.000	0.630	0.830	**0.824**	0.070	0.669	1.016
	Spl	**0.734**	0.000	0.644	0.836	**0.899**	0.378	0.709	1.140
MCML									
	Lin	**0.654**	0.000	0.558	0.767	**0.667**	0.006	0.499	0.890
	Quin	**0.605**	0.000	0.462	0.792	**0.728**	0.281	0.408	1.296
	Spl	**0.669**	0.000	0.561	0.796	**0.710**	0.051	0.504	1.002

Table 3.1: Logistic regression analysis of the NIH-AARP Diet and Health Study for the HEI-2005 total score in Section 3.3. There are five methods considered: (a) moment reconstruction (MR); (b) moment-adjusted imputation (MAI), (c) regression calibration (RC); (d) the food frequency questionnaire (FFQ); and (e) Monte-Carlo maximum likelihood (MCML): the latter three were all done in the original data scale. Within each method the predictor either entered linearly (Lin), via quintiles (Quin) or via a B-spline (Spl). Displayed are the relative risk (RR, in bold face), the p-value, and the lower and upper 95% confidence bounds (L 95% and H 95%, respectively) for the relative risk. The relative risk for the linear and spline analyses were the relative risk for moving from a total score of 45 to a total score of 75, while the relative risk for the quintile analysis was for the quintiles of the usual HEI-2005 total score. The quintile analysis for regression calibration is not included because it is known that categorization induces differential measurement error in regression calibration unless the true risk function is actually a step function of the categories.

imum likelihood all have risks about 10% lower than those estimated by the FFQ, with moment reconstruction between the first three methods and the FFQ. The only anomaly arises in the quintile analysis, where Monte-Carlo maximum likelihood estimates a relative risk 16% smaller than that of the FFQ. The quintile model actually does not fit the data well. This may reflect that had $\mathbf{X}$ been observable, a quintile analysis would have suggested much more attenuation of risk when using the FFQ compared to the linear model.

The results for women are interesting. We do not observe the same phenomenon about the quintile analysis using Monte-Carlo maximum likelihood as was observed in men. The spline model does appear more appropriate than a linear model, and if we look at the spline model results, all the measurement error corrections suggest a large attenuation of risk when using the FFQ. Perhaps of most interest is that when using the FFQ, there is no statistically significant effect of HEI-2005 total score on colorectal cancer when using the FFQ. However, all the measurement error correction methods are different, with p-values ranging from 0.0% to 5.1%. This may seem paradoxical, since the folklore is that measurement error can be ignored when testing null effects, but as discussed in Chapter 10 of Carroll et al. (2006), such folklore is generally true only if there are no covariates measured without error that are also correlated with $\mathbf{X}$. In our case, there are over 20 covariates $\mathbf{Z}$ in the risk model, and, importantly, those covariates are also predictors of $\mathbf{X}$ in the model of Zhang et al. (2011b) as is discussed in Section 3.1.3, and in fact diet composition does depend on the demographic factors making up $\mathbf{Z}$. We believe it is this phenomenon that leads to the change from non-statistical significance to statistical significance in the women.

3.4 Simulation Study

To simulate data that has properties similar to the observed data, several steps are necessary. First, one needs to simulate a calibration data set (usual intake). The calibration data requires specification of model parameters $(\boldsymbol{\Theta}, \boldsymbol{\Sigma}_\zeta, \boldsymbol{\Sigma}_\xi)$, which are estimated in Zhang et al. (2011b). For the purpose of this simulation, we used these aforementioned estimated values as the true model parameters. Given simulated usual intake, one can simulate total score and energy, which are necessary to calculate the risk function associated with colorectal cancer and therefore simulate this outcome. In this simulation study, two different risk functions are considered. Let $H(X) = \{1 + \exp(-X)\}^{-1}$, then

$$\text{pr}(Y = 1|\mathbf{X}, \mathbf{Q}, \mathbf{Z}, \mathbf{U}) = H\left(\gamma_0 + \mathbf{X}^T\gamma_1 + \mathbf{Z}^T\gamma_2\right)$$

The risk functions considered are respectively linear ($\mathbf{X}$ includes total score linearly) and a quintile function ($\mathbf{X}$ includes a step function based on the quintiles of total score). It is then possible to apply the different methods discussed here (MR, MAI, RC, MCML) to use the intake observations with measurement error present to estimate the relative risk associated with an increase in total score. When the specified risk function is linear, both a linear and quintile model are fit, while when the specified risk function is quintile, only a quintile model is fit. In each instance, 500 data sets were generated. In each instance, relative risk (RR) was estimated. Table 3.2 provides a summary of the average RR from the 500 simulations and the standard deviation of the estimated RR.

In summary, none of the methods show serious bias, and moment-adjusted imputation and moment reconstruction are comparable in performance, although moment-adjusted imputation tends to have smaller standard deviation than does moment

reconstruction among females.

Men							
Risk Function	Fit Function		Truth	MR	MAI	MCML	RC
Linear	Linear	RR	0.682	0.681	0.677	0.685	0.685
		10×sd		0.502	0.4810	0.474	0.474
Linear	Quintile	RR	0.682	0.661	0.654	0.635	0.662
		10×sd		0.574	0.554	0.592	0.559
Quintile	Quintile	RR	0.691	0.711	0.704	0.688	0.696
		10×sd		0.624	0.602	0.651	0.599
Women							
Risk Function	Fit Function		Truth	MR	MAI	MCML	RC
Linear	Linear	RR	0.703	0.711	0.707	0.716	0.715
		10×sd		1.058	0.977	1.000	1.000
Linear	Quintile	RR	0.703	0.691	0.680	0.659	0.688
		10×sd		1.273	1.171	1.273	1.206
Quintile	Quintile	RR	0.712	0.748	0.741	0.708	0.715
		10×sd		0.913	0.857	0.891	0.819

Table 3.2: Simulation results of logistic regression for 500 simulated data sets. Displayed are the mean relative risks of moving from the 10^{th} to the 90^{th} percentile of the HEI-2005 total score in the linear analysis and from the 1^{st} to the 5^{th} quintile in the quintile analysis (RR) across the simulation, and 10 $\times$ the standard deviation across the simulations (sd). "Linear" risk function means the disease status is simulated from a logistics model in which the predictor total score enters linearly. "Quintile" risk function means the disease status is simulated from a logistics model which contains the dummy variables of the total score based on quintiles of the total score as predictors. The fit function is "Linear" if the total score enters the model linearly and is "Quintile" if we compare the relative risk of the 1^{st} and 5^{th} quintiles when fitting the model. The methods used are moment reconstruction (MR), moment-adjusted imputation (MAI), Monte Carlo maximum likelihood (MCML), and regression calibration (RC).

4. MEASUREMENT ERROR MODELS WITH ZERO INFLATION AND HARD ZEROS, WITH APPLICATIONS IN NUTRITION WHEN THERE ARE NEVER-CONSUMERS

4.1 Introduction

There is a long history of estimating the distribution of a true variable subject to measurement error. For continuous variables, the literature is enormous, summarized by Carroll et al. (2006) and Buonaccorsi (2010). Analysis of measurement error models of zero-inflated data is more recent, both for estimating the distribution of the true zero-inflated variable in a population, and for disease risk estimation based on the true variable. When only one variable is measured, and it is zero inflated, the literature includes Tooze et al. (2002, 2006) and Kipnis et al. (2009). These are two-part models of non-negative outcomes: the first part models the probability of observing a non-zero outcome on a single observation from an individual, and the second part models the distribution of the observed continuous outcome when it is non-zero. In many problems, however, especially but not limited to nutrition, there are additional variables measured with error that are not subject to excess zeros. Zhang et al. (2011a) considered a measurement error model for one zero-inflated variable and one continuous variable and cast it into a latent variable framework amenable to Markov Chain Monte-Carlo (MCMC) computation. Zhang et al. (2011a) showed that their approach was stable in terms of numerical convergence properties. Zhang et al. (2011b) generalized Zhang et al. (2011a) to the case of multiple continuous and multiple zero-inflated variables, and applied it to dietary patterns research.

There are practical cases, however, when some individuals will always report zero. This occurs, for example, with alcohol consumed from alcoholic beverages,

reported on a daily basis, or with reported red meat consumption. In the former case, the majority of American adults are frequent or occasional consumers of alcoholic beverages, but a fraction never consume alcoholic beverages. Thus, in measurements made on a single occasion, e.g., one day, there are two sources of zeros: the excess zeros caused by episodic consumption, and the hard zeros caused by never consuming alcohol. In most studies, it is not feasible to obtain more than a small number of repeated measurements, e.g. 2-4, on any individual. If all such measurements are zero it is not possible to distinguish whether the person is an episodic consumer or a never consumer. This is what makes the problem so difficult.

The problem of excess and hard zeros was also considered by Kipnis et al. (2009, page 1009) and by Keogh & White (2011), but only for a single variable. They develop a three-part model: the first part for the probability of being a never-consumer, the second part for the probability of consumption on any particular day among those (unknown) individuals who are episodic consumers, and the third part for the continuous measurements on consumption days. They use maximum likelihood with numerical integration over two random effects: Kipnis et al. (2009) use adaptive Gaussian quadrature via the NLMIXED procedure in SAS, and Keogh & White (2011) use Gauss-Hermite quadrature. The former method is computationally slow and prone to converge to a solution on the boundary of the parameter space with a singular Hessian matrix. The latter method has the same issue, and in our experience while it always converges, it frequently announces that the probability of a person being a never-consumer is zero, see Section 3.2.

In this paper, we generalize the model of Kipnis et al. (2009) and Keogh & White (2011) for episodic and never-consumers to allow for a continuous variable to be measured simultaneously, e.g., energy (caloric) intake. In many instances, nutritionists normalize a dietary component by its ratio with the amount of kilo-calories (Guen-

ther et al., 2008a,b), so it is important to be able to model both simultaneously. As in Zhang et al. (2011a,b), estimation in our new model is undertaken by MCMC, which we found to be much more numerically stable than the maximum likelihood approaches of the other authors.

Section 4.2 describes our model and details of prior work. Section 4.3 gives the results of empirical examples, and Section 4.4 gives simulation results for settings similar to our data analysis in Section 4.3. Section 4.5 discusses extensions of the work. Technical details are provided in Appendix A.

4.2 The Model

4.2.1 Review of Kipnis et al. (2009) and Keogh & White (2011)

The original work of Kipnis et al. (2009) and Keogh & White (2011) focused entirely on a single episodically consumed component. Here we briefly review their model. Their observed data have $i = 1, ..., n$ individuals, each with $k = 1, ..., m_i$ repeated measurements, which we here refer to as recalls. In some studies, a large proportion of individuals may have only one measurement. The variables observed are Y_{i1k}, the indicator of whether the food is consumed, and Y_{i2k}, the amount of the food consumed. It is useful to distinguish between covariates $\mathbf{G}_i$ for modeling the probability of being a never-consumer and covariates $\mathbf{X}_i$ for modeling everything else. In Bayesian modeling of binary responses, it is convenient to use a probit model rather than a logistic model (Albert & Chib, 1993), a convention we follow here.

In the probit version of the model of Kipnis et al. (2009) and Keogh & White (2011), they set the probability of being a consumer to be $\Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})$, where $\Phi(\cdot)$ is the standard normal distribution function. Among those who are consumers, they consider random variables $(U_{i1}, U_{i2}) = \text{Normal}(0, \boldsymbol{\Sigma}_u)$, as follows. Given $(\mathbf{X}_i, U_{i1}, U_{i2})$, (Y_{i1k}, Y_{i2k}) are assumed independent of one another and across recalls k, with the

probability of reporting consumption on the k^{th} recall being $\Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1})$.

The reported amount of consumption on a consumption day, as well as reported energy, are positive and quite skewed. A natural and computationally appealing approach used by many in nutrition is to transform such data so that they are more nearly Gaussian. In addition, it generally help numerical stability of the algorithms to standardize the transformed data so that they have mean zero and a fixed variance, which we here take to be 2.0. To do the transformation, both authors use the Box-Cox transformation to a linear mixed model, where the Box-Cox transformation function is $g(x, \lambda) = (x^{\lambda} - 1)/\lambda$ for $\lambda \neq 0$ and $\log(x)$ for $\lambda = 0$. For numerical stability, we also do the standardization, so that we define $S_{i2k}(\lambda_2) = g(Y_{i2k}, \lambda_2)$, define $\mu_2(\lambda_2)$ and $\sigma_2(\lambda_2)$ to be the mean and standard deviation, respectively of the $S_{i2k}(\lambda_2)$ over all observations with $Y_{i2k} > 0$, and finally define $W_{i2k} = \sqrt{2}\{S_{i2k}(\lambda_2) - \mu_2(\lambda_2)\}/\sigma_2(\lambda_2)$. Then their model is that conditionally on $(\mathbf{X}_i, U_{i1}, U_{i2})$, independent of Y_{i1k}, on consumption days $W_{i2k} = \mathrm{Normal}(\mathbf{X}_i\boldsymbol{\beta}_2 + U_{i2}, s_{22})$.

Define $S_i = \sum_{k=1}^{m_i} Y_{i1k}$, the number of recalls in which a non-zero is reported. These authors show that the likelihood for the i^{th} individual given $(\mathbf{X}_i, U_{i1}, U_{i2}, \mathbf{G}_i)$ is

$$\begin{aligned}
\mathcal{L}_{i,\mathrm{obs}} \;=\; & I(S_i = 0)\left[1 - \Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha}) + \Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\left\{1 - \Phi(\mathbf{X}_i^{\mathrm{T}}\beta_1 + U_{i1})\right\}^{m_i}\right] \\
& + I(S_i > 0)\Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\{1 - \Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1})\}^{m_i - S_i}\Phi^{S_i}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1}) \\
& \times s_{22}^{-S_i/2}\textstyle\prod_{k=1}^{m_i}\left[\phi\left\{\left(W_{i2k} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2 - U_{i2}\right)/s_{22}^{1/2}\right\}\right]^{Y_{i1k}}. \qquad (4.1)
\end{aligned}$$

Numerical integration over the distribution of (U_{i1}, U_{i2}) is used to calculate the marginal likelihood. As described in Section 4.1, Kipnis et al. (2009) use the NLMIXED procedure in SAS, which is based on adaptive Gaussian quadrature, while Keogh & White (2011) use Gauss-Hermite quadrature.

4.2.2 Accounting for Energy Intake

The three observed variables are Y_{i1k}, the indicator of whether the food is reported to have been consumed, Y_{i2k}, the reported consumption amount of the food, and Y_{i3k}, the reported amount of energy consumed. For modeling whether a person is a consumer or not ("hard zero"), we consider a latent variable, $\mathcal{N}_i = \text{Normal}(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha}, 1)$, so that a person is a consumer if $\mathcal{N}_i > 0$ and is a never-consumer otherwise. Hence, the marginal probability of being a consumer is $\Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})$. Among consumers, we posit three variables W_{ijk} for $j = 1, 2, 3$, see below for definition, with the properties that

$$W_{ijk} = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_j + U_{ij} + \epsilon_{ijk}, \tag{4.2}$$

with $(U_{i1}, U_{i2}, U_{i3})^{\mathrm{T}} = \text{Normal}(0, \boldsymbol{\Sigma}_u)$, and $(\epsilon_{i1k}, \epsilon_{i2k}, \epsilon_{i3k})^{\mathrm{T}} = \text{Normal}(0, \boldsymbol{\Sigma}_\epsilon)$, where, following Zhang et al. (2011a),

$$\boldsymbol{\Sigma}_\epsilon = \begin{bmatrix} 1 & 0 & \gamma\cos(\theta)s_{33}^{1/2} \\ 0 & s_{22} & \gamma\sin(\theta)(s_{22}s_{33})^{1/2} \\ \gamma\cos(\theta)s_{33}^{1/2} & \gamma\sin(\theta)(s_{22}s_{33})^{1/2} & s_{33} \end{bmatrix}, \tag{4.3}$$

where $\gamma \in (-1, 1)$ and $\theta \in (-\pi, \pi)$.

Responses were transformed via Box-Cox transformations according to the method of Appendix Section A.14: for the episodically consumed component, only positive values were used. We standardized the transformed data, so that for $j = 2, 3$, we write $S_{ijk}(\lambda_j) = g(Y_{ijk}, \lambda_j)$ as a transformed amount consumed. Its inverse transformation is $g^{-1}(S_{ijk}, \lambda_j) = Y_{ijk}$. We define $\mu_2(\lambda_2)$ and $\sigma_2(\lambda_2)$ in the same manner as in Section 4.2.1 and $\mu_3(\lambda_3)$ and $\sigma_3(\lambda_3)$ to be the mean and the standard deviation,

respectively of the $S_{i3k}(\lambda_3)$ over all observations. We then define

$$\begin{aligned} W_{i1k} > 0 \quad &\Longleftrightarrow \quad Y_{i1k} = 1; \\ W_{ijk} \quad &= \quad \sqrt{2}\{S_{ijk}(\lambda_j) - \mu_j(\lambda_j)\}/\sigma_j(\lambda_j), \quad j = 2, 3. \end{aligned}$$

The term W_{i2k}, referring to the consumption amount of the food, is observable if $Y_{i1k} = 1$, i.e., if the food is consumed, but it is latent if $Y_{i1k} = 0$. The term W_{i1k}, referring to whether the food is consumed, is always latent except that we know whether it is positive or not; while the term W_{i3k}, referring to consumption amount of energy, is always observable.

4.2.3 The Complete Data Likelihood Function

The observed data are $(Y_{i1k}, Y_{i2k}, Y_{i3k})$, or equivalently, $(Y_{i1k}, Y_{i1k}W_{i2k}, W_{i3k})$. Define $\widetilde{\mathbf{W}}_{ik} = (W_{i1k}, W_{i2k}, W_{i3k})^{\mathrm{T}}$, $\widetilde{\mathbf{U}}_i = (U_{i1}, U_{i2}, U_{i3})^{\mathrm{T}}$ and $\widetilde{\mathbf{R}}_{ik} = \widetilde{\mathbf{W}}_{ik} - (\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1, ..., \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}} - \widetilde{\mathbf{U}}_i$. The parameters are $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_\epsilon)$. Let $\mathcal{D}_i = I(Y_{i11} = \cdots = Y_{i1m_i} = 0, W_{i11} < 0, ..., W_{i1m_i} < 0)$. Then the likelihood function for the complete data model for person i is

$$\begin{aligned} \mathcal{L}_{i,\text{never}} \quad &\propto \quad |\boldsymbol{\Sigma}_u^{-1}|^{1/2}|\boldsymbol{\Sigma}_\epsilon^{-1}|^{m_i/2} \exp(-\widetilde{\mathbf{U}}_i^{\mathrm{T}}\boldsymbol{\Sigma}_u^{-1}\widetilde{\mathbf{U}}_i/2) \qquad (4.4) \\ &\qquad\qquad \times\phi(\mathcal{N}_i - \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha}) \left\{A_{i1}A_{i3} + A_{i2}A_{i4}\right\}; \\ A_{i1} \quad &= \quad I(\mathcal{N}_i < 0)\mathcal{D}_i; \\ A_{i2} \quad &= \quad I(\mathcal{N}_i > 0); \\ A_{i3} \quad &= \quad (2\pi)^{-3m_i/2}\textstyle\prod_{k=1}^{m_i} \left\{ \dfrac{\exp(-\widetilde{\mathbf{R}}_{ik}^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon^{-1}\widetilde{\mathbf{R}}_{ik}/2)}{1 - \Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1})} \right\}; \\ A_{i4} \quad &= \quad (2\pi)^{-3m_i/2}\textstyle\prod_{k=1}^{m_i} \left\{\exp(-\widetilde{\mathbf{R}}_{ik}^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon^{-1}\widetilde{\mathbf{R}}_{ik}/2)\right\} \\ &\qquad\qquad \times\{Y_{i1k}I(W_{i1k} > 0) + (1 - Y_{i1k})I(W_{i1k} < 0)\}. \end{aligned}$$

The term $1 - \Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1})$ in the denominator of A_{i3} is due to the fact that when $A_{i1} = 1$, the W_{i1k} are truncated normal random variables, truncated from the right at zero.

Remark 1 In Appendix Section A.1, we show that, in the special case that only the episodic component is to be analyzed, when we compute the observed data likelihood function from the complete data likelihood function (4.4), the result coincides and is hence equivalent to the likelihood function of Kipnis et al. (2009) and Keogh & White (2011) and given in (4.1).

4.2.4 Computation

The MCMC calculations, which are a combination of Gibbs (Casella & George, 1992) and Metropolis-Hastings (Chib & Greenberg, 1995) steps, are described in Appendix Sections A.1-A.13. They are similar to those of Zhang et al. (2011a,b), although there are a number of important differences because of the term A_{i3} in (4.4). We program in MATLAB. See Appendix B for the main program.

4.3 Empirical Examples

4.3.1 Overview

We use data from two studies: a subset of women from a case-control study nested within the EPIC-Norfolk study in the U.K. (Day et al., 2001; Bingham et al., 2001) and a subset of the data from the Eating at America's Table Study (EATS) (Subar et al., 2001). In the EPIC-Norfolk subset, each individual has two 7-day food diaries as the instrument, while the EATS data set has four 24-hour recalls as the instrument, both for alcohol from alcoholic beverages and for energy. Both studies have age, body mass index, a food frequency questionnaire for alcohol and a food frequency questionnaire for energy as covariates. The latter two were transformed,

see below, and then all four were centered and standardized to have mean zero and standard deviation one.

Information about the percentages of zeros for the instrument and the FFQ are given in Table 4.1. The percentage of women who report consumption of alcoholic beverages is much higher for EPIC-Norfolk, likely because in that sample, there are 14 days of information in the diet diaries, while the EATS data set is only for 4 days of recalls. Because of this, we see considerable differences between EPIC-Norfolk and EATS in other categories. For example, among those who claim to be non-consumers on the FFQ, four times as many women in EPIC-Norfolk report actual consumption as do the women in EATS. Among those who claim to be consumers on the FFQ, nearly twice as many women report consumption in EPIC-Norfolk as in EATS. Among the women who report no alcoholic beverage consumption on the diary/recall, more than twice as many women in EATS claim to be consumers on the FFQ. All these factors suggest that the estimated percentage of never-consumers will be much less in EPIC-Norfolk than in EATS.

The diary/recall data and the FFQ data were transformed towards normality using the device reported in Appendix Section A.14, and the transformation parameters are given in Table 4.2. After transformation, outliers were removed by deleting those with intake of alcohol more than 2 interquartile ranges above the 75^{th} percentile and those with energy more than 2 interquartile ranges below the 25^{th} percentile or above the 75^{th} percentile. This is a standard device in nutritional epidemiology to remove outliers and leverage points, see also the discussion in Section 4.5.

We did two sets of analyses. In the first, no covariate is used to model for the probability of being a real consumer, so that $\Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha}) = \Phi(\alpha)$. In the second, $\mathbf{G}_i$ consisted of 1.0 for an intercept and the indicator that the FFQ for alcohol equals to 0.

Category	EPIC-Norfolk	EATS Men	EATS Women
Sample size	741	430	497
% who report cons. in one or more diary/recall	78.8%	52.6%	40.6%
Among them, the % who report being consumers on the FFQ	87.8%	93.4%	93.6%
% who report no cons. on any diary/recall	21.2%	47.4%	59.4%
Among them, the % who report being consumers on the FFQ	24.2%	60.3%	56.9%
% whose FFQ reports they are never-consumers	25.6%	22.3%	28.2%
Among them, % whose diaries or recalls report any consumption	37.4%	15.6%	9.3%
% whose FFQ reports they are consumers	74.4%	77.7%	71.8%
Among them, % whose diaries or recalls report any consumption	93.1%	63.2%	52.9%

Table 4.1: Summary statistics for alcohol intake. The EPIC-Norfolk data are based on two 7-day diaries, hence a total of 14 days, while the EATS data are based on 24HR recalls over 4 days.

We analyzed the EPIC-Norfolk data, the EATS data using the first 2 recalls for men and women separately, and the EATS data using all 4 recalls for the men and women separately, for a total of 10 analyses. Our MCMC calculations used as starting values the results from the method of Zhang et al. (2011a), which assumes that everyone is a consumer, and we then used 200,000 MCMC iterations with a burn in of 50,000 iterations, and no convergence problems were noted. The prior distributions used are defined in Appendix Section A.4. Parameter estimates were defined as the posterior means for $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_\epsilon)$ while the posterior mean of $\Phi(\alpha)$ or $n^{-1}\sum_{i=1}^{n}\Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})$ was used as the estimate of the probability of being a consumer. The MCMC steps given in the Appendix A were programmed in MATLAB. The maximum likelihood method with Gauss-Hermite quadrature used 20 quadrature points, and minus the

Variable	EPIC-Norfolk	EATS Men	EATS Women
FFQ, Alcohol	0.05	0.00	0.00
FFQ, Energy	0.37	0.00	0.00
24HR, Alcohol	0.33	0.37	0.44
24HR, Energy	0.73	0.19	0.53

Table 4.2: The Box-Cox transformation parameters used in the data analyses.

loglikelihood was minimized using the MATLAB function "fmincon" with constraints on the parameters to make them finite and with Latin hypercube sampling with 500 grid points to obtain starting values. An R program with optimization using "nlme" produced nearly identical results.

4.3.2 Basic Results for the Percentage of Consumers

Table 4.3 describes the estimates of the percentage of consumers for maximum likelihood with Gauss-Hermite quadrature, which does not use the diary/recall for energy and is based upon (4.1), and for the MCMC method, which does use the diary/recall for energy.

	Covariates	EPIC-Norfolk	EATS, Men		EATS, Women	
	Used for Ever		4	2	4	2
Method	Consume?		Recalls	Recalls	Recalls	Recalls
MLE	No	99.99%	98.62%	99.99%	84.42%	81.75%
	Yes	99.99%	99.99%	99.99%	87.73%	86.01%
MCMC	No	97.34%	92.92%	87.22%	83.41%	83.34%
	Yes	98.43%	95.92%	89.47%	85.97%	85.41%

Table 4.3: The estimated probability of being a consumer for the EPIC-Norfolk and EATS data sets, by method. "Covariates Used for Ever Consume?" indicates whether covariates were used to model the probability of being a consumer.

The most striking result is that in 50% of the analyses, the maximum likelihood estimate of the percentage of consumers is 100%, indicating a lack of real convergence.

In those cases that it does converge, the results are roughly in accord with those of the MCMC analysis. This stability issue with maximum likelihood is to be expected from the work of Kipnis et al. (2009).

However, an unexpected finding is that maximum likelihood is extremely sensitive to the choice of transformation of the alcohol diaries. For EPIC-Norfolk data, we investigated the model with no covariates in the consumer part of the model. We varied the transformation parameter from 0.05 to 0.50. The estimated percentage of consumers was 87%, 88%, 92%, 96%, 100%, 100% and 100% for $\lambda = 0.05$, 0.10, 0.20, 0.25, 0.30, 0.42, and 0.50, respectively. In contrast, the MCMC results varied by less than 1% over this range from the results reported in Table 4.3.

The MCMC results are supplemented with 95% credible intervals in Table 4.4. It is obvious from these tables that in the EATS data, considering all 4 recalls, the percentage of consumers has substantial uncertainty, for men around 10% and for women around 20%. It is an obvious mathematical fact, worth mentioning, that as the percentage of consumers decreases towards 50%, the uncertainty in the estimates of the percentage of consumers will increase. We view these credible interval lengths to be a reflection of the difficulty of the problem.

4.4 Simulations

We simulated data similar to those of Section 4.3. We used the covariate data from our empirical example from the EATS Study of Section 4.3, and used the same transformation of non-zero alcohol amounts and energy. We set $\boldsymbol{\alpha}$ so that there were approximately 90% consumers. There were 200 simulated data sets. In Tables 4.5-4.6 we display the results mimicking men with 4 recalls and without and with covariates in the consumer part of the model. Displayed in these tables are the values of $(\boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\beta})$, the true % of consumers and their estimates.

No Covariates in the Consumer Part of the Model			
	Lower 95^{th}	Posterior mean	Upper 95^{th}
EPIC-Norfolk	94.81%	97.34%	99.14%
EATS, Men, 4 Recalls	86.64%	92.92%	97.66%
EATS, Men, 2 Recalls	78.14%	87.22%	95.74%
EATS, Women, 4 Recalls	74.58%	83.41%	91.64%
EATS, Women, 2 Recalls	71.11%	83.34%	94.03%
With Covariates in the Consumer Part of the Model			
	Lower 95^{th}	Posterior mean	Upper 95^{th}
EPIC-Norfolk	94.65%	98.43%	99.93%
EATS, Men, 4 Recalls	89.29%	95.92%	99.86%
EATS, Men, 2 Recalls	78.97%	89.47%	98.24%
EATS, Women, 4 Recalls	76.69%	85.97%	94.01%
EATS, Women, 2 Recalls	69.93%	85.41%	98.66%

Table 4.4: Posterior analyses of the percentage of consumers, both without and with covariates in the consumer part of the model. Displayed are the posterior mean ("Posterior mean") and the lower ("Lower 95^{th}") and upper ("Upper 95^{th}") 95% credible intervals.

Overall, the simulation study shows that our method does an effective job of estimating the parameters in the model, including estimating the % of consumers. The lengths of the average 95% credible intervals are testimony to the difficulty of this problem with such small number of recalls.

4.5 Discussion

There are a number of generalizations of our work that can be accommodated.

We have concentrated on the case that the episodically consumed food, which is presumed to have never-consumers, has a single accompanying continuous response, in our case energy. In other contexts, the episodically consumed food which is presumed to have never-consumers might have two types of accompanying responses. The first is multivariate continuous responses. The second is a set of semi-continuous responses for foods that are consumed episodically by everyone. Zhang et al. (2011b)

Estimated Percentage of Consumers					
Actual			Estimated, 4 Recalls		
90.0%			87.5% (82.4%,92.0%)		
	$\boldsymbol{\Sigma}_u$			$\widehat{\boldsymbol{\Sigma}}_u$	
0.58	0.03	0.23	0.61 (0.13)	0.10 (0.10)	0.25 (0.06)
0.03	0.52	0.02	0.10 (0.10)	0.52 (0.10)	0.06 (0.07)
0.23	0.02	0.55	0.25 (0.06)	0.06 (0.07)	0.56 (0.06)
	$\boldsymbol{\Sigma}_\epsilon$			$\widehat{\boldsymbol{\Sigma}}_\epsilon$	
1.00	0.00	0.32	1.00	0.00	0.31 (0.05)
0.00	1.02	0.30	0.00	1.05 (0.08)	0.30 (0.06)
0.32	0.30	1.25	0.31 (0.05)	0.30 (0.06)	1.27 (0.05)
	$\boldsymbol{\beta}$			$\widehat{\boldsymbol{\beta}}$	
-0.79	-0.84	0.01	-0.78 (0.08)	-0.89 (0.12)	0.03 (0.05)
0.11	-0.12	-0.12	0.11 (0.07)	-0.12 (0.07)	-0.12 (0.05)
-0.03	0.08	-0.11	-0.02 (0.07)	0.08 (0.08)	-0.11 (0.05)
1.14	1.12	0.00	1.17 (0.10)	1.15 (0.10)	0.02 (0.05)
-0.06	-0.06	0.40	-0.06 (0.07)	-0.06 (0.07)	0.40 (0.04)

Table 4.5: A simulation study of the MCMC method with 200 simulated data sets for EATS men with 4 recalls when no covariates were used in the consumer part of the model. The results shown are the mean estimate over the 200 simulations, and values in parentheses for the parameters are empirical standard deviations. For the estimated percentage of consumers, values in the parentheses represent the average of the 95% credible intervals.

consider this problem but without never-consumers. It appears possible to modify their model and calculations to account for never-consumers in a theoretically straightforward manner, but there are challenges in practice. In particular, the MCMC steps for sampling from the distribution of $\widetilde{\mathbf{U}}_i$ will remain a Metropolis step as in Appendix Section A.11. However, the dimensionality of $\widetilde{\mathbf{U}}_i$ increases as other components are added, and the mixing and convergence of the sampler may be a challenge.

Our model is a measurement error model, and measurement error models intrinsically have a notion of a "true" variable corrupted by measurement error. In such a

Estimated Percentage of Consumers					
Actual			Estimated, 4 Recalls		
91.9%			88.7% (81.2%, 95.7%)		
	$\boldsymbol{\Sigma}_u$			$\widehat{\boldsymbol{\Sigma}}_u$	
0.66	0.07	0.26	0.70 (0.17)	0.13 (0.09)	0.26 (0.08)
0.07	0.51	0.03	0.13 (0.09)	0.52 (0.10)	0.06 (0.07)
0.26	0.03	0.55	0.26 (0.08)	0.06 (0.07)	0.55 (0.06)
	$\boldsymbol{\Sigma}_\epsilon$			$\widehat{\boldsymbol{\Sigma}}_\epsilon$	
1.00	0.00	0.31	1.00	0.00	0.31 (0.05)
0.00	1.02	0.31	0.00	1.04 (0.08)	0.31 (0.06)
0.31	0.31	1.25	0.31 (0.05)	0.31 (0.06)	1.28 (0.05)
	$\boldsymbol{\beta}$			$\widehat{\boldsymbol{\beta}}$	
-0.79	-0.87	0.01	-0.75 (0.09)	-0.92 (0.11)	0.03 (0.05)
0.13	-0.12	-0.11	0.12 (0.07)	-0.12 (0.07)	-0.11 (0.05)
-0.01	0.09	-0.11	-0.01 (0.07)	0.09 (0.07)	-0.11 (0.04)
1.12	1.14	0.00	1.11 (0.09)	1.16 (0.10)	0.01 (0.05)
-0.06	-0.06	0.40	-0.07 (0.07)	-0.06 (0.07)	0.40 (0.04)

Table 4.6: A simulation study of the MCMC method with 200 simulated data sets for EATS men with 4 recalls when the covariate for the probability of being a consumer is the indicator of positive consumption on FFQ. Values in parentheses for the parameters are standard deviations. For the estimated percentage of consumers, values in the parentheses represent the average of the 95% credible intervals.

context, estimating the distribution of the true variable is an important consideration. In this context, following Kipnis et al. (2009) and Keogh & White (2011), it is reasonable to define truth at the individual level as follows. Recall that Y_{i2k} is the reported value of the episodic variable in the original and not the transformed scale, and which may equal zero. Then, by treating Y_{i1k} as unbiased for truth, one may define truth at the individual level as $T_{F,i} = E(Y_{i21}|\mathbf{X}_i, U_{i1}, U_{i2}, \mathbf{G}_i)$. With probability $1 - \Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})$, $T_{F,i} = 0$ or a person is a non-consumer, so the distribution has a point mass at zero. With probability $\Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})$, $T_{F,i} = E(Y_{i21}|\mathbf{X}_i, U_{i1}, U_{i2}, \mathcal{N}_i > 0)$. In the Appendix Sections A.15-A.16, we show how to compute $T_{F,i} = E(Y_{i21}|\mathbf{X}_i, U_{i1}, U_{i2}, \mathcal{N}_i >$

0) and $T_{E,i} = E(Y_{i31}|\mathbf{X}_i, U_{i3}, \mathcal{N}_i > 0)$ using back-transformation. Since we have that $\widetilde{\mathbf{U}}_i = \text{Normal}(0, \boldsymbol{\Sigma}_u)$, Monte-Carlo techniques can be used to estimate the distribution of T across a population (Kipnis et al., 2009). Then the true intake of alcohol can be adjusted by energy as $T_{F,i}/(T_{E,i}/1000)$ and the unit is gram(s) per thousand kilo-calories.

Table 4.7 shows the distribution of usual intake of alcohol, energy and alcohol adjusted by energy among consumers. In the Eating at America's Table Study, men's usual intake of alcohol is about twice of women's on average. Men's energy consumption is about one and a half times of women's. So after adjusting for energy, male consumers' consumption of alcohol is 25% more than female consumers'. Moreover, we see more consumption of alcohol among the women in EPIC-Norfolk data than the women in EATS data. It is due the that fact that the EPIC-Norfolk data are based on 14 days of information (two 7-day diaries) whereas the EATS data are based on 24HR recalls over 4 days.

Finally, it is worth pointing out that the methods, and almost all methods in nutritional epidemiology and measurement error analysis, are not designed to be robust against high leverage outliers. Indeed, in EATS, there was a woman who reported an exceptionally large amount of alcohol intake on the food frequency questionnaire (even in the transformed scale, over 5 interquartile ranges larger than the 75^{th} percentile), yet she claimed no alcohol intake on all 4 recall days. As a result, the mixing of the MCMC sampler was unacceptable, and the estimate of the percentage of consumers was implausibly low. We resorted to the common expedient of removing such outliers according to the method in Section 4.3.1. We caution users of our methods, and measurement error analyses in general, to be aware of the danger of such high leverage outliers.

	Mean	5^{th}	25^{th}	50^{th}	75^{th}	95^{th}
	Alcohol					
EPIC-Norfolk	**8.55**	0.03	1.22	6.26	12.94	25.73
EATS, Women, 4 Recalls	**6.77**	0.01	0.23	2.87	9.54	26.87
EATS, Men, 4 Recalls	**13.05**	0.01	0.41	5.75	19.91	48.71
	Energy					
EPIC-Norfolk	**1758**	1309	1567	1754	1943	2223
EATS, Women, 4 Recalls	**1745**	1128	1465	1722	1999	2436
EATS, Men, 4 Recalls	**2678**	1707	2206	2613	3078	3871
	Ratio					
EPIC-Norfolk	**4.82**	0.02	0.71	3.57	7.27	14.41
EATS, Women, 4 Recalls	**3.96**	0.01	0.14	1.68	5.57	15.62
EATS, Men, 4 Recalls	**4.90**	0.01	0.16	2.21	7.46	18.15

Table 4.7: Estimated distributions of usual intakes among consumers. "Alcohol" means usual intakes of alcohol in gram(s) among consumers. "Energy" means usual intakes of energy in kilo-calories among consumers. "Ratio" means energy-adjusted usual intakes of alcohol among consumers, i.e. amount of alcohol intake / (amount of energy intake / 1000). The unit is gram(s)/(kilo-calories/1000). Displayed are the mean, 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th} percentiles.

5. CONCLUSIONS

The paper first showed that moment reconstruction and moment-adjusted imputation methods can be developed for non-classical measurement error structures, such as this complex, nonlinear Berkson-type measurement error structure. Data analyses and simulation show the promise of these methods in this context.

In the second project, a bivariate measurement error model was developed. The first variable of interest is continuous and positive, e.g. energy intake. The second variable of interest is a mixture of zero and positive measurements with two sources of zeros: episodic zeros and hard zeros. We fit the model using Bayesian methods. The simulations show the new method does well in estimating the parameters in the model, including the % of consumers. Data analyses of the EPIC-Norfolk study in the U.K. and the Eating at America's Table Study show the new method gives more realistic and numerically stable results than the maximum likelihood approach (Keogh & White, 2011). However, we see substantial uncertainty in estimating the % of consumers, especially with only two recalls, as testimony to the difficulty of this problem with such small sample size.

REFERENCES

Albert, J. H. & Chib, S. (1993). Bayesian-analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Bingham, S. A., Welch, A. A., McTaggart, A., Mulligan, A. A., Runswick, S. A., Luben, R., Oakes, S., Khaw, K. T., Wareham, N. & Day, N. E. (2001). Nutritional methods in the european prospective investigation of cancer in norfolk. *Public Health Nutrition* **4**, 847–858.

Buonaccorsi, J. P. (2010). *Measurement error : models, methods, and applications*. Boca Raton, Florida: CRC Press.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective, second edition*. Boca Raton, Florida: Chapman and Hall.

Casella, G. & George, E. I. (1992). Explaining the gibbs sampler. *American Statistician* **46**, 167–174.

Chib, S. & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *American Statistician* **49**, 327–335.

Day, N. E., McKeown, N., Wong, M. Y., Welch, A. & Bingham, S. (2001). Epidemiological assessment of diet: a comparison of a 7-day diary with a food frequency questionnaire using urinary markers of nitrogen, potassium and sodium. *International Journal of Epidemiology* **30**, 309–317.

Freedman, L. S., Feinberg, V., Kipnis, V., Midthune, D. & Carroll, R. J.

(2004). A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics,* **60**, 171–181.

Freedman, L. S., Midthune, D., Carroll, R. J. & Kipnis, V. (2008). A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine,* **27**, 5195–6216.

Guenther, P. M., Reedy, J., Krebs-Smith, S. M. & Reeve, B. B. (2008b). Evaluation of the healthy eating index-2005. *Journal of the American Dietetic Association,* **108**, 1854–1864.

Guenther, P. M., Reedy., J. & S. M. Krebs-Smith, S. M. (2008a). Development of the healthy eating index-2005. *Journal of the American Dietetic Association,* **108**, 1896–1901.

Keogh, R. H. & White, I. R. (2011). Allowing for never and episodic consumers when correcting for error in food record measurements of dietary intake. *Biostatistics* **12**, 624–636.

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., Tooze, J. A., Carroll, R. J. & Freedman, L. S. (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics,* **65**, 1003–1010.

Reedy, J. R., Mitrou, P. N., Krebs-Smith, S. M., Wirfält, E., Flood, A. V., Kipnis, V., Leitzmann, M., Hollenbeck, T. M. A., Schatzkin, A. & Subar, A. F. (2008). Index-based dietary patterns and risk of colorectal

cancer: the nih-aarp diet and health study. *American Journal of Epidemiology,* **168**, 38–48.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing* **5**, 121–125.

Schatzkin, A., Subar, A. F., Thompson, F. E., Harlan, L. C., Tangrea, J., Hollenbeck, A. R., Hurwitz, P. E., Coyle, L., Schussler, N., Michaud, D. S., Freedman, L. S., Brown, C. C., Midthune, D. & Kipnis, V. (2001). Design and serendipity in establishing a large cohort with wide dietary intake distributions: the national institutes of health-aarp diet and health study. *American Journal of Epidemiology,* **154**, 1119–1125.

Subar, A. F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A. & Rosenfeld, S. (2001). Comparative validation of the block, willett, and national cancer institute food frequency questionnaires - the eating at america's table study. *American Journal of Epidemiology* **154**, 1089–1099.

Thomas, L., Stefanski, L. A. & Davidian, M. (2011). A moment-adjusted imputation method for measurement error models. *Biometrics,* **67**, 1461–1470.

Tooze, J. A., Grunwald, G. K. & Jones, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research* **11**, 341–355.

Tooze, J. A., Midthune, D., Dodd, K. W., Freedman, L. S., Krebs-Smith, S. M., Subar, A. F., Guenther, P. M., Carroll, R. J. & Kipnis, V. (2006). A new statistical method for estimating the usual intake of episodically

consumed foods with application to their distribution. *Journal of the American Dietetic Association* **106**, 1575–1587.

Zhang, S., Krebs-Smith, S. M., Midthune, D., Pérez, A., Buckman, D. W., Kipnis, V., Freedman, L. S., Dodd, K. W. & Carroll, R. J. (2011a). Fitting a bivariate measurement error model for episodically consumed dietary components. *International Journal of Biostatistics* **7**, 1–17.

Zhang, S., Midthune, D., Guenther, P. M., Krebs-Smith, S. M., Kipnis, V., Dodd, K. W., Buckman, D. W., Tooze, J. A., Freedman, L. S. & Carroll, R. J. (2011b). A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment. *Annals of Applied Statistics,* **5**, 1456–1487.

APPENDIX A

DETAILS OF CALCULATIONS OF SECTION 4

A.1 Proof of Equivalence With Only One Food

We now show that, in the special case that only the episodic component is to be analyzed, our model is equivalent to that of Kipnis et al. (2009) and Keogh & White (2011). In this case, there is $j = 1, 2$, but no $j = 3$, and we are simply analyzing the usual intake of an episodically consumed dietary component. Also, $\mathbf{\Sigma}_\epsilon$ is a diagonal matrix with entries $(1, s_{22})$. Given (U_{i1}, U_{i2}), and excluding the priors since the referenced papers do not do Bayesian computation, the complete data likelihood function for the i^{th} individual then becomes

$$\begin{aligned}
\mathcal{L}_i &= \phi(\mathcal{N}_i - \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\{I(\mathcal{N}_i < 0)\mathcal{D}_i A_{i5} + I(\mathcal{N}_i > 0)A_{i6}\} \\
&\quad \times s_{22}^{-m_i/2} \prod_{k=1}^{m_i} \phi\{(W_{i2k} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2 - U_{i2})/s_{22}^{1/2}\}; \\
A_{i5} &= \prod_{k=1}^{m_i} \left[\phi(W_{i1k} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 - U_{i1})/\{1 - \Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1})\}\right]; \\
A_{i6} &= \prod_{k=1}^{m_i} \{\phi(W_{i1k} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 - U_{i1})\}\{Y_{i1k} I(W_{i1k} > 0) + (1 - Y_{i1k}) I(W_{i1k} < 0)\}.
\end{aligned}$$

Define $S_i = \sum_{k=1}^{m_i} Y_{i1k}$. To form the observed likelihood function for the observed data given (U_{i1}, U_{i2}), namely Y_{i1k} and $Y_{i1k}W_{i2k}$, we integrate over the latent variables,

namely $\mathcal{N}_i$, (W_{i1k}) and those W_{i2k} for which $Y_{i1k} = 0$. This yields

$$\begin{aligned}\mathcal{L}_{i,\text{obs}} &= I(S_i = 0)\left[1 - \Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha}) + \Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\left\{1 - \Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1})\right\}^{m_i}\right] \\ &\quad + I(S_i > 0)\Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\{1 - \Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1})\}^{m_i - S_i}\Phi^{S_i}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1}) \\ &\quad \times s_{22}^{-S_i/2}\textstyle\prod_{k=1}^{m_i}\left[\phi\{(W_{i2k} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2 - U_{i2})/s_{22}^{1/2}\}\right]^{Y_{i1k}}.\end{aligned}$$

This is exactly the same likelihood function that Kipnis et al. (2009) and Keogh & White (2011) use, with the substitution of the probit for the logistic function.

To see that this argument is true, we need to compute

$$\text{pr}(Y_{i11} = y_1, ..., Y_{i1m_i} = y_{m_i}, Y_{i11}W_{i21} = w_1, ..., Y_{i1m_i}W_{i2m_i} = w_{m_i}|U_{i1}, U_{i2}).$$

This equals zero if for any $k = 1, ..., m_i$, $y_k = 0$ and $w_k \neq 0$ or if $y_k = 1$ and $w_k = 0$. There are then two cases. The first case is when $(y_1 = ... = y_{m_i} = w_1 = ... = w_{m_i} = 0)$. This is simply

$$\begin{aligned}&\text{pr}(Y_{i11} = ... = Y_{i1m_i} = 0|U_{i1}, U_{i2}) \\ &\quad = \text{pr}(\mathcal{N}_i < 0) + \text{pr}(\mathcal{N}_i > 0, Y_{i11} = ... = Y_{i1m_i} = 0|U_{i1}, U_{i2}) \\ &\quad = 1 - \Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha}) + \Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\{1 - \Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1})\}^{m_i},\end{aligned}$$

as claimed. For the other case when $\sum_k y_k > 0$, consider for example

$$\begin{aligned}
&\text{pr}(Y_{i11} = 1, Y_{i12} = ... = Y_{i1m_i} = 0, \\
&\qquad\qquad Y_{i11}W_{i21} = w_1, Y_{i12}W_{i22} = ... = Y_{i1m_i}W_{i2m_i} = 0 | U_{i1}, U_{i2}) \\
&= \text{pr}(\mathcal{N}_i > 0)\text{pr}(Y_{i11} = 1, Y_{i12} = ... = Y_{i1m_i} = 0, \\
&\qquad\qquad Y_{i11}W_{i21} = w_1, Y_{i12}W_{i22} = ... = Y_{i1m_i}W_{i2m_i} = 0 | U_{i1}, U_{i2}, \mathcal{N}_i > 0) \\
&= \Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1})\{1 - \Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1})\}^{m_i - 1} \\
&\qquad\qquad \times s_{22}^{-1/2}\phi\{(w_1 - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2 - U_{i2})/s_{22}^{1/2}\}.
\end{aligned}$$

The other cases are similar.

A.2 Initial Details

Define $N = \sum_{i=1}^{n} m_i$. Below, by "rest", we mean all the observable data, latent variables and parameters other than the one in question.

All covariates are pre-standardized to have mean zero and variance one except the intercept term.

Zhang et al. (2011a) point out that $\boldsymbol{\Sigma}_\epsilon$ is a full rank matrix with determinant $\det(\boldsymbol{\Sigma}_\epsilon) = s_{22}s_{33}(1 - \gamma^2)$.

A.3 The Truncated Normal Distribution

We use the notation $\text{TN}_+(\mu, \sigma, c)$ for a normal random variable with mean μ, standard deviation σ , truncated from the left at c, and $\text{TN}_-(\mu, \sigma, c)$ is truncated from the right at c.

A.4 Prior Distributions and Definitions

Because the data were standardized, and following the implementation of Zhang et al. (2011a), we used the following conventions. The results from the method of

Zhang et al. (2011a) are used as starting value of $\mathbf{\Sigma}_u$ and $\boldsymbol{\beta}$. Starting values of r and θ both equal 0.00 and starting values of s_{22} and s_{33} both equal 1.00. Starting values of $\boldsymbol{\alpha}$ are the same as their prior means.

- The priors for all $\boldsymbol{\beta}_j$ are normal random variables with mean zero and diagonal covariance matrix $\mathbf{\Omega}_{\beta,j}$ with variance 10.00.

- When there are no covariates, the prior for α is Normal($\alpha_{\text{prior}} = 0.8416, \Omega_\alpha = 0.40^2$), which reflects a prior mean of being a consumer of roughly 80%. For $\boldsymbol{\alpha}$, when there are covariates, the prior distribution is normal with mean zero and diagonal covariance matrix with standard deviation 1.00.

- The prior mean $\mathbf{\Omega}$ for $\mathbf{\Sigma}_u$ is exchangeable with diagonal entries all equal to 1.0 and correlations 0.50. There was $m_u = 5$ degrees of freedom in the inverse Wishart prior. Thus, with the dimensionality of $\mathbf{\Sigma}_u$ $p_{\text{dim}} = 3$, the prior density is

$$\begin{aligned} f_{\text{IW}}(\mathbf{\Sigma}_u, \mathbf{\Omega}_u, m_u, p_{\text{dim}}) &= (m_u - p_{\text{dim}} - 1)^{-m_u/2} |\mathbf{\Omega}_u|^{m_u/2} \\ &\quad \times |\mathbf{\Sigma}_u^{-1}|^{(m_u + p_{\text{dim}} + 1)/2} \exp[-\text{trace}\{(m_u - p_{\text{dim}} - 1)\mathbf{\Omega}_u \mathbf{\Sigma}_u^{-1}/2\}]. \end{aligned}$$

  This density has mean $\mathbf{\Omega}_u$ for any m_u.

- The priors for s_{22} and s_{33} are Uniform[0,3]. This range is reasonable because of the standardization.

- The priors for (γ, θ) are uniform on their range.

- Denote $\phi(x)$ as the standard normal density function.

A.5 Complete Conditionals for $\mathcal{N}_i$

If $\sum_{k=1}^{m_i} Y_{i1k} > 0$, then $[\mathcal{N}_i|\text{rest}, \sum_{k=1}^{m_i} Y_{i1k} > 0] \propto \phi(\mathcal{N}_i - \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})I(\mathcal{N}_i > 0)$, and hence

$$[\mathcal{N}_i|\text{rest}, \textstyle\sum_{k=1}^{m_i} Y_{i1k} > 0] = \text{TN}_+(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha}, 1, 0) = \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha} + \text{TN}_+(0, 1, -\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha}).$$

If $\sum_{k=1}^{m_i} Y_{i1k} = 0$, then $[\mathcal{N}_i|\text{rest}, \sum_{k=1}^{m_i} Y_{i1k} = 0] \propto \phi(\mathcal{N}_i - \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\left\{A_{i3}I(\mathcal{N}_i < 0) + A_{i4}I(\mathcal{N}_i > 0)\right\}$, and hence

$$[\mathcal{N}_i|\text{rest}, \textstyle\sum_{k=1}^{m_i} Y_{i1k} = 0] = \dfrac{\phi(\mathcal{N}_i - \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\left\{A_{i3}I(\mathcal{N}_i < 0) + A_{i4}I(\mathcal{N}_i > 0)\right\}}{A_{i3}\{1 - \Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\} + A_{i4}\Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})}.$$

This means that when $\sum_{k=1}^{m_i} Y_{i1k} = 0$, $[\mathcal{N}_i|\text{rest}, \sum_{k=1}^{m_i} Y_{i1k} = 0]$ is a mixture of truncated normal random variables, which can be simulated using the algorithm of Robert (1995). Define

$$\begin{aligned} p_i &= A_{i3}\{1 - \Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\}/[A_{i3}\{1 - \Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\} + A_{i4}\Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})] \\ &= \left[1 + \frac{\Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})\{1 - \Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1})\}^{m_i}}{1 - \Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})}\right]^{-1}. \end{aligned}$$

Then, using the truncated normal notation defined in Section A.3,

$$\begin{aligned} [\mathcal{N}_i|\text{rest}, \textstyle\sum_{k=1}^{m_i} Y_{i1k} = 0] &= \text{TN}_-(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha}, 1, 0) \text{ with probability } p_i; \\ &= \text{TN}_+(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha}, 1, 0) \text{ with probability } 1 - p_i, \end{aligned}$$

or

$$\begin{aligned}[\mathcal{N}_i|\text{rest}, \textstyle\sum_{k=1}^{m_i} Y_{i1k} = 0] &= \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha} - \mathrm{TN}_+(0, 1, \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha}) \text{ with probability } p_i; \\ &= \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha} + \mathrm{TN}_+(0, 1, -\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha}) \text{ with probability } 1 - p_i.\end{aligned}$$

When $\sum_{k=1}^{m_i} Y_{i1k} = 0$, p_i increases with increasing m_i, which is intuitively appealing, because the more observed zero intakes, the greater the confidence that the subject is a never-consumer.

A.6 Complete Conditionals for $\boldsymbol{\alpha}$

Except for irrelevant constants

$$\begin{aligned}[\boldsymbol{\alpha}|\text{rest}] &= \exp\{-(1/2)\boldsymbol{\alpha}^{\mathrm{T}}\mathcal{C}_2^{-1}\boldsymbol{\alpha} + \mathcal{C}_1^{\mathrm{T}}\boldsymbol{\alpha}\}; \\ \mathcal{C}_2 &= (\textstyle\sum_{i=1}^{n} \mathbf{G}_i\mathbf{G}_i^{\mathrm{T}} + \boldsymbol{\Omega}_\alpha^{-1})^{-1}; \\ \mathcal{C}_1 &= \boldsymbol{\Omega}_\alpha^{-1}\boldsymbol{\alpha}_{\text{prior}} + \textstyle\sum_{i=1}^{n} \mathcal{N}_i\mathbf{G}_i.\end{aligned}$$

This means that

$$[\boldsymbol{\alpha}|\text{rest}] = \text{Normal}(\mathcal{C}_2\mathcal{C}_1, \mathcal{C}_2).$$

A.7 Complete Conditionals for $(\gamma, \theta, s_{22}, s_{33})$

The complete conditionals for $(\gamma, \theta, s_{22}, s_{33})$ do not have an explicit form, so we use a Metropolis-Hastings within Gibbs sampler to generate them in turn. Since $\boldsymbol{\Sigma}_\epsilon$ is determined by γ, θ, s_{22} and s_{33}, we write it as $\boldsymbol{\Sigma}_\epsilon{}^{-1} \equiv f(\gamma, \theta, s_{22}, s_{33})$. Also, current values are γ_t, θ_t, $s_{22,t}$ and $s_{33,t}$.

Generation of γ. For convenience, we set γ to be discrete with 41 equally-spaced values on its range. The candidate value y is selected randomly from γ_t and its

two nearest neighbors. The candidate value y is accepted with probability $p(\gamma_t, y)$, $p(\gamma_t, y) = \min\{1, g(y)/g(\gamma_t)\}$, where

$$\begin{aligned} g(y) &\propto (1-y^2)^{-N/2} \\ &\quad \times \exp\left[-(1/2)\textstyle\sum_{i=1}^{n}\sum_{k=1}^{m_i}\{\widetilde{\mathbf{W}}_{ik} - (\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1, ..., \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}} - \widetilde{\mathbf{U}}_i\}^{\mathrm{T}} f(y, \theta_t, s_{22,t}, s_{33,t})\{\bullet\}\right], \end{aligned}$$

where $\{\bullet\}$ means that the term before $f(\cdot)$ is transposed and substituted. If the candidate y is accepted, then $\gamma_{t+1} = y$. Otherwise, $\gamma_{t+1} = \gamma_t$.

Generation of θ. This is done exactly as for γ, except now

$$g(y) \propto \exp\left[-(1/2)\textstyle\sum_{i=1}^{n}\sum_{k=1}^{m_i}\{\widetilde{\mathbf{W}}_{ik} - (\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1, ..., \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}} - \widetilde{\mathbf{U}}_i\}^{\mathrm{T}} f(\gamma_{t+1}, y, s_{22,t}, s_{33,t})\{\bullet\}\right].$$

If the candidate y is accepted, then $\theta_{t+1} = y$. Otherwise, $\theta_{t+1} = \theta_t$.

Generation of s_{22}. A candidate value y is generated from the Uniform distribution of length 0.4 with mean $s_{22,t}$: y = Uniform[$s_{22,t}$ - 0.2, $s_{22,t}$ + 0.2]. The candidate value y is accepted with probability $p(s_{22,t}, y)$, where

$$\begin{aligned} p(s_{22,t}, y) &= \min\left\{(1, g(y)I_{[0,3]}(y)/g(s_{22,t})\right\}; \\ g(y) &\propto y^{-N/2}\exp\left[-(1/2)\textstyle\sum_{i=1}^{n}\sum_{k=1}^{m_i}\{\widetilde{\mathbf{W}}_{ik} - (\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1, ..., \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}} - \widetilde{\mathbf{U}}_i\}^{\mathrm{T}}\right. \\ &\qquad \left. \times f(\gamma_{t+1}, \theta_{t+1}, y, s_{33,t})\{\bullet\}\right]. \end{aligned}$$

If the candidate is accepted, then $s_{22,t+1} = y$. Otherwise, $s_{22,t+1} = s_{22,t}$.

Generation of s_{33}. This is the same as that for s_{22}, except now

$$\begin{aligned} p(s_{33,t}, y) &= \min\left\{1, g(y) I_{[0,3]}(y)/g(s_{33,t})\right\}; \\ g(y) &\propto y^{-N/2} \exp\big[-(1/2)\textstyle\sum_{i=1}^{n}\sum_{k=1}^{m_i}\{\widetilde{\mathbf{W}}_{ik} - (\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1, ..., \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}} - \widetilde{\mathbf{U}}_i\}^{\mathrm{T}} \\ &\qquad \times f(\gamma_{t+1}, \theta_{t+1}, s_{22,t+1}, y)\{\bullet\}\big]. \end{aligned}$$

If the candidate is accepted, then $s_{33,t+1} = y$. Otherwise, $s_{33,t+1} = s_{33,t}$.

A.8 Complete Conditional for $\boldsymbol{\Sigma}_u$

Define $B_1 = m_u + n$ and $B_2 = \{(m_u - p_{\mathrm{dim}} - 1)\boldsymbol{\Omega}_u + \sum_{i=1}^{n}\widetilde{\mathbf{U}}_i\widetilde{\mathbf{U}}_i^{\mathrm{T}}\}$. Then it is easily seen that

$$[\boldsymbol{\Sigma}_u|\mathrm{rest}] = \mathrm{IW}(\boldsymbol{\Sigma}_u, B_2, B_1, p_{\mathrm{dim}}).$$

A.9 Complete Conditionals for $(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$

Let the (ℓ, p) element of $\boldsymbol{\Sigma}_{\epsilon}^{-\mathbf{1}}$ be $\sigma_{\epsilon}^{\ell,p}$. Then, for $j = 2, 3$, except for irrelevant constants,

$$\begin{aligned} \log\left[\boldsymbol{\beta}_j|\mathrm{rest}\right] &= -(1/2)\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\Omega}_{\beta,j}^{-1}\boldsymbol{\beta}_j \\ &\quad -(1/2)\textstyle\sum_{i=1}^{n}\sum_{k=1}^{m_i}(W_{ijk} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_j - U_{ij})^2\sigma_{\epsilon}^{jj} \\ &\quad -\textstyle\sum_{i=1}^{n}\sum_{k=1}^{m_i}\sum_{\ell\neq j}\sigma_{\epsilon}^{j\ell}(W_{ijk} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_j - U_{ij})(W_{i\ell k} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell - U_{i\ell}) \\ &= \mathcal{C}_1^{\mathrm{T}}\boldsymbol{\beta}_j - (1/2)\boldsymbol{\beta}_j^{\mathrm{T}}\mathcal{C}_2^{-1}\boldsymbol{\beta}_j \end{aligned}$$

where

$$\begin{aligned}
\mathcal{C}_2 &= (\boldsymbol{\Omega}_{\beta,j}^{-1} + \textstyle\sum_{i=1}^{n} m_i \sigma_\epsilon^{jj} \mathbf{X}_i \mathbf{X}_i^{\mathrm{T}})^{-1}; \\
\mathcal{C}_1 &= \textstyle\sum_{i=1}^{n} \sum_{k=1}^{m_i} \sigma_\epsilon^{jj} \mathbf{X}_i (W_{ijk} - U_{ij}) \\
&\quad + \textstyle\sum_{i=1}^{n} \sum_{k=1}^{m_i} \sum_{\ell \neq j} \sigma_\epsilon^{j\ell} (W_{i\ell k} - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_\ell - U_{i\ell}) \mathbf{X}_i.
\end{aligned}$$

This implies that for $j = 2, 3$, $\left[\boldsymbol{\beta}_j | \text{rest}\right] = \text{Normal}(\mathcal{C}_2 \mathcal{C}_1, \mathcal{C}_2)$,

A.10 Complete Conditionals for $\boldsymbol{\beta}_1$

Define

$$\begin{aligned}
\mathcal{C}_2 &= \{\boldsymbol{\Omega}_{\beta,1}^{-1} + \textstyle\sum_{i=1}^{n} m_i \sigma_\epsilon^{11} \mathbf{X}_i \mathbf{X}_i^{\mathrm{T}}\}^{-1}; \\
\mathcal{C}_1 &= \textstyle\sum_{i=1}^{n} \sum_{k=1}^{m_i} \sigma_\epsilon^{11} \mathbf{X}_i (W_{i1k} - U_{i1}) \\
&\quad + \textstyle\sum_{i=1}^{n} \sum_{k=1}^{m_i} \sum_{\ell \neq 1} \sigma_\epsilon^{1\ell} (W_{i\ell k} - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_\ell - U_{i\ell}) \mathbf{X}_i; \\
c_1(\boldsymbol{\beta}_1) &= \exp(\mathcal{C}_1^{\mathrm{T}} \boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^{\mathrm{T}} \mathcal{C}_2^{-1} \boldsymbol{\beta}_1 / 2); \\
c_2(\boldsymbol{\beta}_1) &= \prod_{i=1}^{n} \{1 - \Phi(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_1 + U_{i1})\}^{-m_i I(\mathcal{N}_i < 0)}.
\end{aligned}$$

Then the complete conditional density

$$[\boldsymbol{\beta}_1 | \text{rest}] \propto h(\boldsymbol{\beta}_1) = c_1(\boldsymbol{\beta}_1) c_2(\boldsymbol{\beta}_1).$$

A closed form for the posterior of $\boldsymbol{\beta}_1$ is not available and we use the random walk Metropolis proposal,

$$q(\boldsymbol{\beta}_{1,\text{cand}} | \boldsymbol{\beta}_{1,\text{curr}}) = \text{Normal}(\boldsymbol{\beta}_{1,\text{curr}}, \boldsymbol{\Sigma}).$$

The Metropolis ratio is

$$\frac{c_1(\boldsymbol{\beta}_{1,\text{cand}})c_2(\boldsymbol{\beta}_{1,\text{cand}})}{c_1(\boldsymbol{\beta}_{1,\text{curr}})c_2(\boldsymbol{\beta}_{1,\text{curr}})}.$$

A reasonable choice is to set $\boldsymbol{\Sigma} = M\mathcal{C}_2$. In our calculations, we use $M = 2$ since this results in good mixing.

A.11 Complete Conditionals for $\widetilde{\mathbf{U}}_i$

Define

$$\begin{aligned}
\mathcal{C}_2 &= (\boldsymbol{\Sigma}_u^{-1} + m_i{\boldsymbol{\Sigma}_\epsilon}^{-1})^{-1};\\
\mathcal{C}_1 &= \textstyle\sum_{k=1}^{m_i}{\boldsymbol{\Sigma}_\epsilon}^{-1}\{\widetilde{\mathbf{W}}_{ik} - (\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1, ..., \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}}\};\\
c_1(\widetilde{\mathbf{U}}_i) &= \exp(\mathcal{C}_1^{\mathrm{T}}\widetilde{\mathbf{U}}_i - \widetilde{\mathbf{U}}_i^{\mathrm{T}}\mathcal{C}_2^{-1}\widetilde{\mathbf{U}}_i/2);\\
c_2(U_i) &= \{1 - \Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1})\}^{-m_i I(\mathcal{N}_i<0)}.
\end{aligned}$$

The likelihood function $[\widetilde{\mathbf{U}}_i|\text{rest}] \propto c_1(\widetilde{\mathbf{U}}_i)c_2(\widetilde{\mathbf{U}}_i)$. Of course, $c_1(\widetilde{\mathbf{U}}_i)$ is proportional to the density of a Normal$(\mathcal{C}_2\mathcal{C}_1, \mathcal{C}_2)$. Thus, if $\mathcal{N}_i > 0$, $\widetilde{\mathbf{U}}_i$ = Normal$(\mathcal{C}_2\mathcal{C}_1, \mathcal{C}_2)$. If $\mathcal{N}_i < 0$, then we have to do a Metropolis step, and we use the candidate density Normal$(\mathcal{C}_2\mathcal{C}_1, \mathcal{C}_2)$. If $\widetilde{\mathbf{U}}_{i,\text{curr}}$ and $\widetilde{\mathbf{U}}_{i,\text{cand}}$ are the current and candidate values, respectively, then we accept the candidate with probability

$$\min\{1, c_2(\widetilde{\mathbf{U}}_{i,\text{cand}})/c_2(\widetilde{\mathbf{U}}_{i,\text{curr}})\}.$$

An alternative when $\mathcal{N}_i < 0$ is to make the candidate density Normal$(\widetilde{\mathbf{U}}_{i,\text{curr}}, 2\mathcal{C}_2)$, in which case we accept the candidate with probability

$$\min[1, c_1(\widetilde{\mathbf{U}}_{i,\text{cand}})c_2(\widetilde{\mathbf{U}}_{i,\text{cand}})/\{c_1(\widetilde{\mathbf{U}}_{i,\text{curr}})c_2(\widetilde{\mathbf{U}}_{i,\text{curr}})\}].$$

A.12 Complete Conditionals for W_{i1k}

Here we do the complete conditional for W_{i1k}. Except for irrelevant constants,

$$\begin{aligned}
\log[W_{i1k}|\text{rest}] &= (-1/2)\{Y_{i1k}I(W_{i1k}>0)+(1-Y_{i1k})I(W_{i1k}<0)\} \\
&\quad \times(W_{i1k}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1-U_{i1},...,W_{i3k}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_3-U_{i3})\boldsymbol{\Sigma}_\epsilon{}^{-1}(\bullet) \\
&= -\{Y_{i1k}I(W_{i1k}>0)+(1-Y_{i1k})I(W_{i1k}<0)\} \\
&\qquad \times\big\{(1/2)\sigma_\epsilon^{11}(W_{i1k}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1-U_{i1})^2 \\
&\qquad\quad +\textstyle\sum_{j\neq 1}\sigma_\epsilon^{1j}(W_{i1k}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1-U_{i1})(W_{ijk}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_j-U_{ij})\big\} \\
&= \{Y_{i1k}I(W_{i1k}>0)+(1-Y_{i1k})I(W_{i1k}<0)\} \\
&\qquad \times\left\{\mathcal{C}_1W_{i1k}-(1/2)W_{i1k}^2\mathcal{C}_2^{-1}\right\},
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{C}_2 &= 1/(\sigma_\epsilon^{11}) \\
\mathcal{C}_1 &= \sigma_\epsilon^{11}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1+U_{i1})-\textstyle\sum_{j\neq 1}\sigma_\epsilon^{1j}(W_{ijk}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_j-U_{ij}).
\end{aligned}$$

Using the truncated normal notation defined in Section A.3, it follows that with $\mu=\mathcal{C}_2\mathcal{C}_1$ and $\sigma=\mathcal{C}_2^{1/2}$,

$$\begin{aligned}
[W_{i1k}|\text{rest}] &= Y_{i1k}\mathrm{TN}_+(\mu,\sigma,0)+(1-Y_{i1k})\mathrm{TN}_-(\mu,\sigma,0) \\
&= \mu+Y_{i1k}\mathrm{TN}_+(0,\sigma,-\mu)+(1-Y_{i1k})\mathrm{TN}_-(0,\sigma,-\mu) \\
&= \mu+Y_{i1k}\mathrm{TN}_+(0,\sigma,-\mu)-(1-Y_{i1k})\mathrm{TN}_+(0,\sigma,\mu) \\
&= \mu+\sigma\{Y_{i1k}\mathrm{TN}_+(0,1,-\mu/\sigma)-(1-Y_{i1k})\mathrm{TN}_+(0,1,\mu/\sigma)\}.
\end{aligned}$$

A.13 Complete Conditionals for W_{i2k} When it is Not Observed

The variable W_{i2k} is not observed when $Y_{i1k} = 0$, or, equivalently, when $W_{i1k} < 0$. Except for irrelevant constants,

$$\begin{aligned} \log\left[W_{i2k}|\text{rest}\right] &= -(1/2)\sum_j \sum_\ell \sigma_\epsilon^{j\ell}(W_{ijk} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_j - U_{ij})(W_{i\ell k} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell - U_{i\ell}) \\ &= -(1/2)W_{i2k}^2\mathcal{C}_2^{-1} + \mathcal{C}_1 W_{i2k} \end{aligned}$$

where

$$\begin{aligned} \mathcal{C}_2 &= 1/(\sigma_\epsilon^{22}); \\ \mathcal{C}_1 &= \sigma_\epsilon^{22}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2 + U_{i2}) - \sum_{\ell \neq 2} \sigma_\epsilon^{2\ell}(W_{i\ell k} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell - U_{i\ell}). \end{aligned}$$

Therefore,

$$[W_{i2k}|\text{rest}] = W_{i2k}Y_{i1k} + (1 - Y_{i1k})\text{Normal}(\mathcal{C}_2\mathcal{C}_1, \mathcal{C}_2).$$

A.14 Transformation Estimation

The algorithm is as follows. Define $g(x, \lambda) = (x^\lambda - 1)/\lambda$ if $\lambda \neq 0$ and set $g(x, \lambda) = \log(x)$ if $\lambda = 0$.

For any variable, take the non-zero data and compute the first through the 99^{th} percentiles as $(x_1, ..., x_{99})$. Then for each lambda, define $Y(\lambda) = \{g(x_1, \lambda), ..., g(x_{99}, \lambda)\}^{\mathrm{T}}$. Form the Blom scores, $B = (b_1, ..., b_{99})$ as $b_i = \Phi^{-1}\{(i - 3/8)/(n_x + 1/4)\}, i = 1, \ldots, 99$, where $\Phi(\cdot)$ is the normal distribution function and $n_x = 99$. Then define $G(\lambda)$ to be the R^2 in the regression of $Y(\lambda)$ on B, and choose λ to maximize this R^2.

A.15 Distribution of Usual Intake

Let $a_{\min,\mathrm{F}}$ be 1/2 the minimum value of Y_{i2k} for those (i,k) such that $Y_{i2k} > 0$. This is defined because of worries that the back-transformation might lead to ridiculously negative intakes on consumption days. Similarly, let $a_{\min,\mathrm{E}}$ be 1/2 the minimum value of Y_{i3k} over (i,k).

With probability $\Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})$, a person is a consumer, while with probability $1-\Phi(\mathbf{G}_i^{\mathrm{T}}\boldsymbol{\alpha})$, a person is a never-consumer. The percentage of never-consumers in the population is estimated as $1-n^{-1}\sum_{i=1}^{n}\Phi(\mathbf{G}_i^{\mathrm{T}}\widehat{\boldsymbol{\alpha}})$.

Among consumers, at the individual level, the chance that the person consumes on a given day is $\Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1+U_{i1})$. Recall that the amount reported to be consumed on a consumption day is Y_{i2k} while that of energy is Y_{i3k}, that $S_{ijk}(\lambda_j)=g(Y_{ijk},\lambda_j)$ and that $W_{ijk}=\sqrt{2}\{S_{ijk}(\lambda_j)-\mu_j(\lambda_j)\}/\sigma_j(\lambda_j)$. Hence,

$$\begin{aligned} Y_{ijk} &= g^{-1}\{\mu_j(\lambda_j)+\sigma_j(\lambda_j)W_{ijk}/\sqrt{2},\lambda_j\} \\ &= g^{-1}\left\{\mu_j(\lambda_j)+\sigma_j(\lambda_j)(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_j+U_{ij}+\epsilon_{ijk})/\sqrt{2},\lambda_j\right\}. \end{aligned}$$

The average amount of the episodically consumed food on consumption days is

$$Q_F(\mathbf{X}_i,\boldsymbol{\beta}_2,s_{22},U_{i2})=E\left[g^{-1}\left\{\mu_2(\lambda_2)+\sigma_2(\lambda_2)(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2+U_{i2}+\epsilon_{i2k})/\sqrt{2},\lambda_2\right\}|U_{i2}\right] \quad \text{(A.1)}$$

True usual intake for consumers is defined as

$$T_F(\mathbf{X}_i,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,s_{22},U_{i1},U_{i2})=\max\{a_{\min,\mathrm{F}},Q_F(\mathbf{X}_i,\boldsymbol{\beta}_2,s_{22},U_{i2})\}\Phi(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1+U_{i1}). \quad \text{(A.2)}$$

Similarly, define

$$Q_E(\mathbf{X}_i, \boldsymbol{\beta}_3, s_{33}, U_{i3}) = E\left[g^{-1}\left\{\mu_3(\lambda_3) + \sigma_3(\lambda_3)(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_3 + U_{i3} + \epsilon_{i3k})/\sqrt{2}, \lambda_3\right\}|U_{i3}\right] \quad \text{(A.3)}$$

so that true usual intake of energy is

$$T_E(\mathbf{X}_i, \boldsymbol{\beta}_3, s_{33}, U_{i3}) = \max\{a_{\min,\mathrm{E}}, Q_E(\mathbf{X}_i, \boldsymbol{\beta}_3, s_{33}, U_{i3})\}.$$

A.16 Computation of Back-transformed Expectation

The function (A.1) can be approximated by Gauss-Hermite quadrature, which tries to approximate integrals of the form

$$\int f(x)\exp(-x^2)dx \approx \sum_{\ell} w_\ell f(x_\ell),$$

where the w_ℓ are the weights and the x_ℓ are the abscissas. In nutritional epidemiology, it is traditional to use 9-point Gauss-Hermite quadrature, with the abscissas and weights

$$\begin{aligned}(x_1, ..., x_9) &= (-2.1, -1.3, -0.8, -0.5, 0.00, 0.5, 0.8, 1.3, 2.1);\\ (w_1, ..., w_9) &= (0.063345, 0.080255, 0.070458, 0.159698, 0.252489,\\ &\qquad 0.159698, 0.070458, 0.080255, 0.063345).\end{aligned}$$

It can be verified that the integrals are exact with $f(x) = x^k$ for $k = 0, 1, 2, 3, 4, 5$, and of course for all odd functions.

Remembering that for $j = 2, 3$ $\epsilon_{ijk} = \mathrm{Normal}(0, s_{jj})$, with a change of variable

we can rewrite (A.1) as

$$Q_F(\mathbf{X}_i, \boldsymbol{\beta}_2, s_{22}, U_{i2}) \tag{A.4}$$
$$= \frac{1}{\sqrt{\pi}} \int \exp(-x^2) g^{-1} \left\{ \mu_2(\lambda_2) + \sigma_2(\lambda_2)(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_2 + U_{i2} + x s_{22}^{1/2} \sqrt{2})/\sqrt{2}, \lambda_2 \right\} dx.$$

When $\lambda = 0$, $g^{-1}(v, \lambda) = \exp(v)$, and so computation of (A.4) is simple and direct. When $\lambda \neq 0$, $g^{-1}(v, \lambda) = (1 + \lambda v)^{1/\lambda}$, which only makes sense if $1 + \lambda v > 0$. Thus, we have to make sure that in the computation with abscissas x_ℓ that

$$1 + \lambda_2 \left\{ \mu_2(\lambda_2) + \sigma_2(\lambda_2)(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_2 + U_{i2} + x_\ell s_{22}^{1/2} \sqrt{2})/\sqrt{2} \right\} > 0.$$

When $\lambda_2 \neq 0$, the quadrature approximation to (A.1) then becomes

$$Q_F(\mathbf{X}_i, \boldsymbol{\beta}_2, s_{22}, U_{i2}) \tag{A.5}$$
$$\approx \frac{1}{\sqrt{\pi}} \sum_{\ell=1}^{9} w_\ell \max \left[0, 1 + \lambda_2 \left\{ \mu_2(\lambda_2) + \sigma_2(\lambda_2)(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_2 + U_{i2} + x_\ell s_{22}^{1/2} \sqrt{2})/\sqrt{2} \right\} \right]^{1/\lambda_2}.$$

Similarly,

$$Q_E(\mathbf{X}_i, \boldsymbol{\beta}_3, s_{33}, U_{i3}) \tag{A.6}$$
$$\approx \frac{1}{\sqrt{\pi}} \sum_{\ell=1}^{9} w_\ell \max \left[0, 1 + \lambda_3 \left\{ \mu_3(\lambda_3) + \sigma_3(\lambda_3)(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_3 + U_{i3} + x_\ell s_{33}^{1/2} \sqrt{2})/\sqrt{2} \right\} \right]^{1/\lambda_3}.$$

APPENDIX B

MATLAB CODE OF SECTION 4

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% This is the MCMC analysis for a single food plus energy in the three-part
% model. It can analyze 2 - 4 recalls for EATS data and 2 recalls for
% Norfolk data, both with and without covariates in the ever consumers
% model.
% Call:
%     aarp_setprior_Sigmau
%     backtransform_20130925
%     boxcoxtrans_20130925
%     formGofSigmae
%     gen_truncated_normals
%     gen_Wtildei_1foodplusenergy
%     generate_latex_report
%     generate_usual_intake
%     ginverse
%     load_lambda_neverconsumers
%     load_names
%     make_percentiles_without_weight
%     process_data
%     update_beta1_with_prior_mean_random_walk
%     update_beta1_with_prior_mean
%     update_beta2_with_prior_mean
%     update_beta3_with_prior_mean
%     update_iSigmau
%     update_Utildei
%     update_Ni_with_covariates
%     updated_parameter_r
%     updated_parameter_s22
%     updated_parameter_s33
%     updated_parameter_theta
% Last revised by Rubin Wei <rubin@stat.tamu.edu> on 10/14/2013.
% MATLAB Version: 8.1.0.604 (R2013a)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Initial Setup
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Tell the program 1) which data set (EATS or Norfolk), 2) which food
% (Alcohol, Fish or Deep Yellow Vegetables), 3) which gender (men or
% women), 4) how many recalls, 5) if you want to include covariates in
% alpha and 6) how many realizations of usual intake to generate
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
clear
dataset_ind = 1;            % Which data set: 1: EATS; 2: Norfolk
data_ind    = 1;            % Which food 1: Alcohol; 2: Fish;
                            % 3: Deep Yellow Vegetables
thesex      = 1;            % Sex: 0: Men;  1: Women
mmi         = 4;            % Number of recalls, integer,
                              % for Norfolk data it has to be 2;
                              % for EATS data it should be lareger
                              % than 1 but no more than 4.
```

```
with_covariates_ind = 3;    % What to include as covariates in the
                            % ever consumer model
                              % 0: a column of ones.
                              % 1: a column of ones, the FFQ, and
                              % the indicator that the FFQ=0.
                              % 2: a column of ones and the FFQ.
                              % 3: a column of ones and the indicator
                              % that the FFQ=0.
n_gen = 4;                  % Number of realizations of usual intake to
                              % generate. Must be a positive integer,
                              % or 0 if no realizations to be generated.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Set up number of MCMC steps, number of burn-in period, number of
% thinning.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
nMCMC     = 200000; % Number of MCMC iterations
nburn     = 50000;  % Size of the burn-in
nthin     = 50;     % Size of thinning
ndist     = 200;    % average of the last ndist MCMC steps (after thinning)
                    % to get the cdf of usual intake
if ndist * nthin > nMCMC
    error('Please decrease ndist or increase nMCMC')
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% To make the MCMC work well, we experimented different combinations of
% settings. for 1) prior mean and starting value of beta;
% 2) proposal distribution of beta_1; 3) prior mean and starting value
% of Sigmau; 4) prior mean, prior variance and starting value of alpha.
% Among the combinations, the following combination works well and is
% our final choice. It not necessary to change it.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
beta_start_ind = 9999;      % the starting value of beta
                              % 9999: episodically;
                              % all other number: a 5*3 matrix with the
                              % number as each cell
beta_prior_mean_ind  = 0;   % the prior mean for beta
                              % 9999: episodically (estimated by Saijuan's
                              % code)
                              % 0: regular (a 5*3 matrix with all
                              % elements = 0)
rw_ind = 1;                 % do you want to use the random walk proposal
                            % for beta_1
                              % 1: yes, use randon walk proposal
                              % Normal(\beta_{1,\curr}, \C_2 /M)
                              % 0: no, use Normal(\C_2 \C_1, \C_2 /M)
update_beta1_var_ind = 0.5; % the variance for updating beta1, i.e. C1
                            % in section A.9, this is the M in
                            % Normal(\beta_{1,\curr}, \C_2 /M)
                            % and Normal(\C_2 \C_1, \C_2 /M)
Sigmau_start_ind = 1;       % the starting value of Sigmau
                              % 1: episodically (estimated by Saijuan's
                              % code)
                              % 2: regular (a 3*3 matrix with diagonal
                              % elements = 1 and off-diagonal
                              % elements = 0.5)
Sigmau_prior_mean_ind = 2;  % the prior mean for Sigmau
                            % 1: episodically (estimated by Saijuan's code)
                            % 2: regular (a 3*3 matrix with diagonal
                            % elements = 1 and off-diagonal elements = 0.5)
                            % 3: half of regular
consumer_percent_prior_mean = ((with_covariates_ind == 0) * 0.8 ...
```

```
                                    + (with_covariates_ind ~= 0) * 0.5);
                                % the prior mean of the percentage of consumers
                                  % 0.5 for with covariates in alpha;
                                  % 0.8 for without covariates in alpha
consumer_percent_start = ((with_covariates_ind == 0) * 0.8 ...
                                    + (with_covariates_ind ~= 0) * 0.5);
                                %starting value for the percentage of consumers
                                  % 0.5 for with covariates in alpha;
                                  % 0.8 for without covariates in alpha
alpha_prior_sd = ((with_covariates_ind == 0) * 0.4 ...
                                    + (with_covariates_ind ~= 0) * 1);
                                % prior standard deviation for alpha
                                  % 1 for with covariates in alpha;
                                  % 0.4 for without covariates in alpha
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Initialized random seed and set up some other stuff
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
ndim     = 3; % the number of dimensions, here = 3 for indicator, amount
              % and energy
beta1_accept_count = 0; % count how many times beta1 moves
% initialize random seed
myseed = 6309021;
format compact;
rand('state',myseed);
randn('state',myseed);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% The following code is for the purpose of creating output folder name and
% some outputs
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
if thesex == 0;
    sex_name = 'Male'; % 'MEN'
elseif thesex == 1;
    sex_name = 'Female'; % 'WOMEN'
else
    error('Sex must be 0 for male and 1 for female.')
end
[beta_start_method, beta_prior_mean_method, update_beta1_var_method, ...
    Sigmau_start_method, Sigmau_prior_mean_method, dataset_name, ...
    with_covariates_method, data_name]...
    = load_names(beta_start_ind, beta_prior_mean_ind, rw_ind, ...
    update_beta1_var_ind, Sigmau_start_ind, Sigmau_prior_mean_ind, ...
    dataset_ind, with_covariates_ind, data_ind);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Creat folder and diary
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% This is the folder where the results will be saved to.
output_folder = ['Output_', dataset_name, '_', data_name, '_', ...
    num2str(mmi), '_recalls_', with_covariates_method, '_', ...
    datestr(now, 'mm_dd_yyyy')];
% This tells the combinations of setting.
output_folder_complete = ['Output_', dataset_name, '_', data_name, '_', ...
    num2str(mmi), '_recalls_', with_covariates_method, '_MCMC_', ...
    num2str(nMCMC), '_beta_', beta_start_method,'_prior_', ...
    beta_prior_mean_method,'_', update_beta1_var_method, ...
    'Sigmau_starting_', Sigmau_start_method, '_prior_mean_', ...
    Sigmau_prior_mean_method, '_percent_prior_', ...
    num2str(consumer_percent_prior_mean), '_start_', ...
    num2str(consumer_percent_start), '_alpha_sd_', ...
    num2str(alpha_prior_sd), '_', datestr(now, 'mm_dd_yyyy')];
% Make the directory
mkdir(output_folder)
```

```
% Diary
diary([output_folder,'/AARP_',sex_name,'_',data_name,'_Diary.txt']);
tic
disp(['Current time is: ', datestr(now, 'yyyy-mm-dd HH:MM:SS')])
disp(['Seed number is: ', num2str(myseed)])
disp(['Folder name is:  ', output_folder])
disp(['The combination of setting is:  ', output_folder_complete])
disp(['Gender is: ', sex_name])
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Get the Box-Cox transformation parameter for recall food
% (lambda_rec_food), recall energy (lambda_rec_energy), ffq food
% (lambda_ffq_food), ffq energy (lambda_ffq_energy)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
[lambda_rec_food, lambda_rec_energy, lambda_ffq_food, lambda_ffq_energy]...
    =  load_lambda_neverconsumers(thesex, dataset_ind, data_ind);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Load the data
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
if dataset_ind == 1 % EATS data
    zz = csvread(['../EATS_3.23/EATS_',data_name, '_', sex_name,...
        '_4Recalls_Outlier_Cleared.csv']);
    Age_ind = 1;
    bmi_ind = 3;
    ffq_food_ind = 4;
    ffq_energy_ind = 5;
    if mmi == 4
    % use all 4 recalls
        rec_food_ind = [6:9];
        rec_energy_ind = [10:13];
    elseif mmi == 2
    % use the first 2 recalls
        rec_food_ind = [6:7];
        rec_energy_ind = [10:11];
    else
    error('Please specify which columns are recall food and recall energy')
    end
elseif dataset_ind == 2 % Norfolk data
    zz = csvread('../Norfolk/Norfolk_Outlier_Cleared.csv');
    if thesex ~= 1
        error('The Norfolk data are all women, please set thesex = 1.')
    end
    if mmi ~= 2
        error('The Norfolk data only has 2 recalls, please set mmi = 2. ')
    end
    Age_ind = 5;
    bmi_ind = 6;
    ffq_food_ind = 7;
    ffq_energy_ind = 8;
    rec_food_ind = [1:2];
    rec_energy_ind = [3:4];
end
n = size(zz, 1);   % sample size
[Wistar, Wi2, Wi3, Age, bmi, ffq_food, ffq_energy, nointake_ffq, ...
    didconsume, a0_food, a0_energy, mumu, sigsig, mu_e, sig_e] = ...
    process_data(zz, n, mmi, Age_ind, bmi_ind, ffq_food_ind, ...
    ffq_energy_ind, rec_food_ind, rec_energy_ind, lambda_ffq_food, ...
    lambda_ffq_energy, lambda_rec_food, lambda_rec_energy);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Save half of the minimum positive food value, half of minimum energy
% value, and Box-Cox transformation parameters.
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
save([output_folder,'/AARP_',sex_name,'_a0_',data_name,'.mat'],'-mat', ...
    'a0_food');
save([output_folder,'/AARP_',sex_name,'_a0_energy.mat'],'-mat', ...
    'a0_energy');
save([output_folder,'/AARP_',sex_name,'_lambda_REC_',data_name,'.mat'],...
    '-mat', 'lambda_rec_food');
save([output_folder,'/AARP_',sex_name,'_lambda_REC_Energy.mat'], ...
    '-mat', 'lambda_rec_energy');
save([output_folder,'/AARP_',sex_name,'_lambda_FFQ_',data_name,'.mat'],...
    '-mat', 'lambda_ffq_food');
save([output_folder,'/AARP_',sex_name,'_lambda_FFQ_Energy.mat'],...
    '-mat', 'lambda_ffq_energy');
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Set up the design matrices in the consmption model.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Xtildei = repmat([ones(n,1) Age bmi ffq_food ffq_energy],[1,1,3]);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Set up the design matrices in the evewr consumer model.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
switch with_covariates_ind
    case 1
        GGalpha = [ones(n,1) ffq_food nointake_ffq];
    case 2
        GGalpha = [ones(n,1) ffq_food];
    case 3
        GGalpha = [ones(n,1) nointake_ffq];
    case 0
        GGalpha = ones(n,1);
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Set the MCMC parameters
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
disp('Set the MCMC parameters');
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Set the prior and starting value for alpha
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
prior_alpha_mean      = norminv(consumer_percent_prior_mean,0,1).* ...
                            ones(size(GGalpha,2),1);
prior_alpha_cov       = (alpha_prior_sd .^ 2).*eye(size(GGalpha,2));
alpha_start           = norminv(consumer_percent_start,0,1).* ...
                            ones(size(GGalpha,2),1);
alpha                 = alpha_start;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Set the prior and starting value for beta
% Good starting values for the beta parameters and their covariance
% matrices are available by other means. I ran the consumption program and
% the amount program to get these values.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
beta_temp=load(['../Saijuan_code/Output_Saijuan_episodically_2Recalls/',...
    dataset_name, '_', data_name,'/AARP_',sex_name,'_',data_name, ...
    '_beta_postmean.mat'], '-mat',  'beta_postmean');
beta_temp = beta_temp.beta_postmean;
if beta_prior_mean_ind  == 9999
    prior_beta_mean  = beta_temp;
elseif beta_prior_mean_ind == 0
    prior_beta_mean       = zeros(5,3);
end
prior_beta_cov = repmat((10 .* eye(5)), [1,1,3]);
if beta_start_ind == 9999
```

```
    beta_start = beta_temp;
else
    beta_start = beta_start_ind .* ones(size(prior_beta_mean));
end
beta                  = beta_start;
clear beta_start
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Set the prior and starting value for Sigmae
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
r     = 0;
theta = 0;
s22   = 1;
s33   = 1;
R     =  [1            0            r*cos(theta)
          0            1            r*sin(theta)
          r*cos(theta) r*sin(theta) 1           ];
A = diag([1 sqrt(s22) sqrt(s33)], 0);
Sigmae       = A'*R*A;
iSigmae      = inv(Sigmae);
Sigmae_start = Sigmae;
prior_Sigmae_doff = 5;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Set the prior and starting values for Sigmau
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
[Sigmau_temp_regular, prior_Sigmau_doff] = aarp_setprior_Sigmau;
Sigmau_temp_episodically = ...
    load(['../Saijuan_code/Output_Saijuan_episodically_2Recalls/', ...
    dataset_name, '_', data_name,'/AARP_',sex_name,'_',data_name, ...
    '_Sigmau_postmean.mat'], '-mat',  'Sigmau_postmean');
Sigmau_temp_episodically = Sigmau_temp_episodically.Sigmau_postmean;
switch Sigmau_prior_mean_ind
    case 1; prior_Sigmau_mean = Sigmau_temp_episodically ;
    case 2; prior_Sigmau_mean = Sigmau_temp_regular ;
    case 3; prior_Sigmau_mean = Sigmau_temp_regular ./ 2;
    otherwise
        error('Sigmau_prior_mean_ind not recognized')
end

switch Sigmau_start_ind
    case 1; Sigmau_start = Sigmau_temp_episodically ;
    case 2; Sigmau_start = Sigmau_temp_regular ;
    otherwise
        error('Sigmau_start_ind not recognized')
end

 Sigmau    = Sigmau_start;
iSigmau    = inv(Sigmau);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Initialize a few things
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Set starting values for the Utildei
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Utildei = randn(n,size(Sigmau,2)) * sqrtm(Sigmau);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Get starting values for the W_{ijk}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
WtildeiS(:,2,:)  = Wi2;
WtildeiS(:,3,:)  = Wi3;
WtildeiS(:,1,:)  = abs(repmat(squeeze(Xtildei(:,:,1)) * beta(:,1) ...
```

```
    + Utildei(:,1), [1,mmi]) + randn(n,mmi));
WtildeiS(:,1,:)  = (squeeze(WtildeiS(:,1,:)) .* Wistar) ...
    - (squeeze(WtildeiS(:,1,:)) .* (1 - Wistar));

numgen           = 20;
Wtildeinew       = gen_Wtildei_1foodplusenergy(WtildeiS,beta,Xtildei, ...
                                   Utildei,n,iSigmae,Wistar,mmi, numgen);
Wtildei          = Wtildeinew;
Wtildei_start = Wtildei;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Initialize the MCMC traces
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
r_trace      = zeros(nMCMC,1);
theta_trace  = zeros(nMCMC,1);
s22_trace    = zeros(nMCMC,1);
s33_trace    = zeros(nMCMC,1);
Sigmae_trace = zeros(3,3,nMCMC);
Sigmau_trace = zeros(3,3,nMCMC);
beta_trace   = zeros(5,3,nMCMC);
alpha_trace  = zeros(nMCMC,size(GGalpha,2));
never_trace  = zeros(nMCMC,1);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Initialize the matrix, whichi is used calculte distribution of food and
% energy
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
usual_intake_food_trace = NaN(n, ndist);
usual_intake_energy_trace = NaN(n, ndist);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% MCMC
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
disp('Start the MCMC');
for jjMCMC = 1:nMCMC;
    if(rem(jjMCMC,500) ==0)
        disp(['iteration = ', num2str(jjMCMC)])
    end;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Update Ni. You create this for everyone.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    [Ni,ppi] = update_Ni_with_covariates(Xtildei,beta,Utildei,alpha,...
        GGalpha,n,mmi,didconsume);
    isnever  = (Ni < 0); % Indicator of a never-consumer
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Update alpha. In the following, the complete conditional for alpha is
% that is a truncated normal from the left at alpha_min, but with mean
% (cc2 * cc1) and variance cc2.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    xx       = squeeze(Xtildei(:,:,1));
    mmnn     = size(xx,2);
    cc1      = (inv(prior_alpha_cov)*prior_alpha_mean) + GGalpha'*Ni;
    cc2      = inv(GGalpha'*GGalpha + inv(prior_alpha_cov));
    mujj     = cc2 * cc1;
    sijj     = sqrtm(cc2);
    alpha    = mujj + sijj*randn(size(GGalpha,2),1);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Update W1 and W2
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    numgen     = 5;
    Wtildeinew = gen_Wtildei_1foodplusenergy(Wtildei,beta,Xtildei,...
        Utildei,n,iSigmae,Wistar,mmi,numgen);
    Wtildei    = Wtildeinew;
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Calculate W-XB-U
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    tt = zeros(n,ndim);
    for jj = 1:ndim;
     tt(:,jj) = (Xtildei(:,:,jj) * beta(:,jj)) + Utildei(:,jj) ;
    end;
    qq = Wtildei(:,:,:) -  repmat(tt,[1,1,mmi]);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Update iSigmae
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    rnew     = updated_parameter_r(r,theta,s22,s33,qq,mmi,n);
    r        = rnew;
    thetanew = updated_parameter_theta(r,theta,s22,s33,qq,mmi);
    theta    = thetanew;
    s22new   = updated_parameter_s22(r,theta,s22,s33,qq,mmi,n);
    s22      = s22new;
    s33new   = updated_parameter_s33(r,theta,s22,s33,qq,mmi,n);
    s33      = s33new;

    R     =  [1            0            r*cos(theta)
              0            1            r*sin(theta)
              r*cos(theta) r*sin(theta) 1           ];
    A = diag([1 sqrt(s22) sqrt(s33)], 0);
    Sigmae       = A'*R*A;
    iSigmae      = inv(Sigmae);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Update iSigmaU
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    [Sigmau_new,iSigmau_new] = update_iSigmau(Sigmau,prior_Sigmau_doff, ...
                             prior_Sigmau_mean,Utildei,n);
    Sigmau  = Sigmau_new;
    iSigmau = iSigmau_new;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Update Utildei. This is done in two steps. In the first step, we generate
% it assuming that everyone is a consumer. In the second step, those who
% are never consumers, i.e., Ni < 0, have their values updated by a
% Metropolis step.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    Utildei_new = update_Utildei(Utildei,beta,Wtildei,iSigmae, ...
                             Ni,isnever,didconsume,Xtildei,mmi,iSigmau,n);
    Utildei     = Utildei_new;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Update beta1 using a Metropolis Step.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    if rw_ind == 1
        beta1 = update_beta1_with_prior_mean_random_walk(Xtildei,mmi, ...
            prior_beta_mean, prior_beta_cov,beta,Wtildei, Utildei, ...
            iSigmae,isnever,update_beta1_var_ind);
    else
        beta1 = update_beta1_with_prior_mean(Xtildei,mmi,...
            prior_beta_mean,prior_beta_cov,beta,Wtildei, Utildei,...
            iSigmae,isnever,update_beta1_var_ind);
    end
    % count if beta1 moves
    beta1_accept_count = beta1_accept_count + (1 -all(beta1 == beta(:,1)));
    beta(:,1)  = beta1;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Update beta2. This does not need a Metropolis step
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    beta2 = update_beta2_with_prior_mean(Xtildei,mmi, prior_beta_mean, ...
```

```
        prior_beta_cov,beta,Wtildei, Utildei,iSigmae);
    beta(:,2)  = beta2;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Update beta2. This does not need a Metropolis step
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    beta3 = update_beta3_with_prior_mean(Xtildei,mmi, prior_beta_mean, ...
        prior_beta_cov,beta,Wtildei, Utildei,iSigmae);
    beta(:,3)  = beta3;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Store results
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    Sigmae_trace(:,:,jjMCMC) = Sigmae;
    Sigmau_trace(:,:,jjMCMC) = Sigmau;
    beta_trace(:,:,jjMCMC)   = beta;
    r_trace(jjMCMC,1)        = r;
    theta_trace(jjMCMC,1)    = theta;
    s22_trace(jjMCMC,1)      = s22;
    s33_trace(jjMCMC,1)      = s33;
    alpha_trace(jjMCMC,:)    = alpha;
    never_trace(jjMCMC,1)    = 1 - sum(normcdf(GGalpha*alpha))./n;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Compute distribution of usual intake.
% Suppose we have finished an MCMC step. In this step, we know who are
% non-consumers (N_i < 0), and who are consumers (N_i > 0).
% Use Gauss-Hermite quadrature method to approximate the Q_F,
% which is average amount of food on consumption day for consumers
% (equations A.5 in section A.16). This is done using the
% backtransform_20130925 function.
% Then plug it in to compute the usual intake for consumers (equation A.2
% in section A.15).
% Do this for about 200 MCMC steps near the end, with thinning of 50.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    if any((nMCMC - (ndist - 1) * nthin):nthin:nMCMC == jjMCMC)
        uuindex = (Ni > 0);
        nindex = sum(uuindex);
        Utildei = randn(n,size(Sigmau,2)) * sqrtm(Sigmau);
        temp = backtransform_20130925(lambda_rec_food, ...
            Xtildei(uuindex,:,2), beta(:,2), sqrt(Sigmae(2,2)), mumu, ...
            sigsig, Utildei(uuindex,2), nindex);
        temp = max(a0_food, temp) .* normcdf(Xtildei(uuindex,:,1) * ...
            beta(:,1) +Utildei(uuindex,1));% get usual intake (n*1), eq A.2
        usual_intake_food = temp;
        temp = backtransform_20130925(lambda_rec_energy, ...
            Xtildei(uuindex,:,3), beta(:,3), sqrt(Sigmae(3,3)), mu_e, ...
            sig_e, Utildei(uuindex,3), nindex);
        usual_intake_energy = max(a0_energy, temp);
        % store the results for this run
        usual_intake_food_trace(:,((jjMCMC - (nMCMC - (ndist - 1)* ...
            nthin))/nthin + 1)) = [usual_intake_food; NaN(n - nindex, 1)];
        usual_intake_energy_trace(:,((jjMCMC - (nMCMC - (ndist - 1)* ...
            nthin))/nthin + 1)) = [usual_intake_energy;NaN(n - nindex, 1)];
    end
end;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% end of MCMC
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
disp('MCMC completed');

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Thinning, burn-in, compute posterior mean, standard deviation,
% credible interval, Compute distribution of usual intake, save results
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Thinning and save the traces (after thinning, but before burn-in)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
  nn             = size(r_trace,1);
  mm             = floor((nn +eps) ./ nthin);
  thin_index     = nthin:nthin:(mm*nthin);
  alpha_thin_trace   =  alpha_trace(thin_index,:);
  never_thin_trace   =  never_trace(thin_index,1);
      r_thin_trace   =      r_trace(thin_index,1);
  theta_thin_trace   =  theta_trace(thin_index,1);
    s22_thin_trace   =    s22_trace(thin_index,1);
    s33_thin_trace   =    s33_trace(thin_index,1);
 Sigmae_thin_trace = Sigmae_trace(:,:,thin_index);
 Sigmau_thin_trace = Sigmau_trace(:,:,thin_index);
   beta_thin_trace =   beta_trace(:,:,thin_index);

save([output_folder,'/AARP_',sex_name,'_',data_name,...
    '_Sigmau_trace.mat'],'-mat','Sigmau_thin_trace');
save([output_folder,'/AARP_',sex_name,'_',data_name,...
    '_Sigmae_trace.mat'],'-mat','Sigmae_thin_trace');
save([output_folder,'/AARP_',sex_name,'_',data_name,...
    '_beta_trace.mat'],'-mat',  'beta_thin_trace');
save([output_folder,'/AARP_',sex_name,'_',data_name,...
    '_alpha_trace.mat'],'-mat',  'alpha_thin_trace');
save([output_folder,'/AARP_',sex_name,'_',data_name,...
    '_never_trace.mat'],'-mat',  'never_thin_trace');
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Get rid of the burn-in
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
  mmburn             = floor((nburn +eps) ./ nthin);
  alpha_thin_trace   =  alpha_thin_trace((mmburn+1):mm,:);
  never_thin_trace   =  never_thin_trace((mmburn+1):mm,1);
      r_thin_trace   =      r_thin_trace((mmburn+1):mm,1);
  theta_thin_trace   =  theta_thin_trace((mmburn+1):mm,1);
    s22_thin_trace   =    s22_thin_trace((mmburn+1):mm,1);
    s33_thin_trace   =    s33_thin_trace((mmburn+1):mm,1);
 Sigmae_thin_trace   = Sigmae_thin_trace(:,:,(mmburn+1):mm);
 Sigmau_thin_trace   = Sigmau_thin_trace(:,:,(mmburn+1):mm);
   beta_thin_trace   =   beta_thin_trace(:,:,(mmburn+1):mm);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Compute and save the posterior means, standard deviation and
% 95% credible interval
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
alpha_postmean      =  mean(alpha_thin_trace,1);
never_postmean      =  mean(never_thin_trace);
beta_postmean       =  mean(beta_thin_trace, 3);
Sigmau_postmean     =  mean(Sigmau_thin_trace, 3);
Sigmae_postmean     =  mean(Sigmae_thin_trace,3);

alpha_postsd        = std(alpha_thin_trace,0,1);
never_postsd        = std(never_thin_trace);
beta_postsd         = std(beta_thin_trace,0,3);
Sigmau_postsd       = std(Sigmau_thin_trace,0,3);
Sigmae_postsd       = std(Sigmae_thin_trace,0,3);

alpha_ci            = quantile(alpha_thin_trace,[0.025, 0.975],1);
never_ci            = quantile(never_thin_trace,[0.025, 0.975]);
beta_ci             = quantile(beta_thin_trace,[0.025, 0.975],3);
Sigmau_ci           = quantile(Sigmau_thin_trace,[0.025, 0.975],3);
Sigmae_ci           = quantile(Sigmae_thin_trace,[0.025, 0.975],3);
```

```
% Computer the correlation matrix for Sigmau
Corru_postmean = NaN(size(Sigmau_postmean));
for iicorr = 1:size(Sigmau_postmean, 1)
    for jjcorr = 1:size(Sigmau_postmean, 2)
        Corru_postmean(iicorr,jjcorr) = Sigmau_postmean(iicorr,jjcorr)/...
            sqrt(Sigmau_postmean(iicorr,iicorr) * ...
            Sigmau_postmean(jjcorr,jjcorr));
    end
end
% Computer the correlation matrix for Sigmae
Corre_postmean = NaN(size(Sigmae_postmean));
for iicorr = 1:size(Sigmae_postmean, 1)
    for jjcorr = 1:size(Sigmae_postmean, 2)
        Corre_postmean(iicorr,jjcorr) = Sigmae_postmean(iicorr,jjcorr)/...
            sqrt(Sigmae_postmean(iicorr,iicorr) * ...
            Sigmae_postmean(jjcorr,jjcorr));
    end
end

 save([output_folder,'/AARP_',sex_name,'_',data_name,...
     '_Sigmau_postmean.mat'],'-mat','Sigmau_postmean');
 save([output_folder,'/AARP_',sex_name,'_',data_name,...
     '_Sigmae_postmean.mat'],'-mat','Sigmae_postmean');
 save([output_folder,'/AARP_',sex_name,'_',data_name,...
     '_beta_postmean.mat'],'-mat',  'beta_postmean');
 save([output_folder,'/AARP_',sex_name,'_',data_name,...
     '_alpha_postmean.mat'],'-mat',  'alpha_postmean');
 save([output_folder,'/AARP_',sex_name,'_',data_name,...
     '_never_postmean.mat'],'-mat',  'never_postmean');

 save([output_folder,'/AARP_',sex_name,'_',data_name,...
     '_Sigmau_postsd.mat'],'-mat','Sigmau_postsd');
 save([output_folder,'/AARP_',sex_name,'_',data_name,...
     '_Sigmae_postsd.mat'],'-mat','Sigmae_postsd');
 save([output_folder,'/AARP_',sex_name,'_',data_name,...
     '_beta_postsd.mat'],'-mat',  'beta_postsd');
 save([output_folder,'/AARP_',sex_name,'_',data_name,...
     '_alpha_postsd.mat'],'-mat',  'alpha_postsd');
 save([output_folder,'/AARP_',sex_name,'_',data_name,...
     '_never_postsd.mat'],'-mat',  'never_postsd');

 save([output_folder,'/AARP_',sex_name,'_',data_name,'_Sigmau_ci.mat'],...
     '-mat','Sigmau_ci');
 save([output_folder,'/AARP_',sex_name,'_',data_name,'_Sigmae_ci.mat'],...
     '-mat','Sigmae_ci');
 save([output_folder,'/AARP_',sex_name,'_',data_name,'_beta_ci.mat'],...
     '-mat',  'beta_ci');
 save([output_folder,'/AARP_',sex_name,'_',data_name,'_alpha_ci.mat'],...
     '-mat',  'alpha_ci');
 save([output_folder,'/AARP_',sex_name,'_',data_name,'_never_ci.mat'],...
     '-mat',  'never_ci');

 save([output_folder,'/AARP_',sex_name,'_',data_name,'_GGalpha.mat'],...
     '-mat',  'GGalpha');
 save([output_folder,'/AARP_',sex_name,'_',data_name,'_Xtildei.mat'],...
     '-mat',  'Xtildei');
 save([output_folder,'/AARP_',sex_name,'_',data_name,'_mu_energy.mat'],...
     '-mat',  'mu_e');
 save([output_folder,'/AARP_',sex_name,'_',data_name,'_sig_energy.mat'],...
     '-mat',  'sig_e');
```

```
 save([output_folder,'/AARP_',sex_name,'_',data_name,'_mu_',data_name,...
     '.mat'], '-mat', 'mumu');
 save([output_folder,'/AARP_',sex_name,'_',data_name,'_sig_',data_name,...
     '.mat'], '-mat', 'sigsig');

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Compute distribution of usual intake of food, energy and
% food/(energy/1000).
% Compute the cdf of usual intake for consumers on a fine grid.
% We have an estimate of the cdf for consumers,
% which can be inverted to get the percentiles.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
usual_intake_ratio_trace = 1000 .* usual_intake_food_trace ./ ...
                                  usual_intake_energy_trace;
food_p_mat = linspace(0,1,501)';
aa = usual_intake_food_trace(~isnan(usual_intake_food_trace));
food_distribution = make_percentiles_without_weight(aa, food_p_mat);
mu_ui_food = mean(aa);
sig_ui_food = std(aa);
%plot(0.01:0.01:0.99, food_distribution)

energy_p_mat = linspace(0,1,501)';
aa = usual_intake_energy_trace(~isnan(usual_intake_energy_trace));
energy_distribution = make_percentiles_without_weight(aa, energy_p_mat);
mu_ui_energy = mean(aa);
sig_ui_energy = std(aa);
%plot(0.01:0.01:0.99, energy_distribution)

ratio_p_mat = linspace(0,1,501)';
aa = usual_intake_ratio_trace(~isnan(usual_intake_ratio_trace));
ratio_distribution = make_percentiles_without_weight(aa, ratio_p_mat);
mu_ui_ratio = mean(aa);
sig_ui_ratio = std(aa);
%plot(0.01:0.01:0.99, ratio_distribution)

if dataset_ind == 1
    adj_factor = 1;
elseif dataset_ind == 2
    adj_factor = 4.184;
end
ui_percentile_ind = [5, 10, 25, 50, 75, 90, 95];
temp1 = food_distribution(ui_percentile_ind)';
temp2 = energy_distribution(ui_percentile_ind)'./adj_factor;
temp3 = ratio_distribution(ui_percentile_ind)'.*adj_factor;
disp(['The mean, 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles '])
disp(['of usual intake food, energy and ratio among consumers are:'])
disp(['&', num2str(mu_ui_food, '%.4f'),'&',num2str(temp1(1),'%.4f'),'&',...
    num2str(temp1(2), '%.4f'),'&',num2str(temp1(3), '%.4f'),'&', ...
    num2str(temp1(4), '%.4f'),'&',num2str(temp1(5), '%.4f'),'&', ...
    num2str(temp1(6), '%.4f'),'&',num2str(temp1(7), '%.4f'),'\\*[-.60em]'])
disp(['&', num2str(mu_ui_energy/adj_factor, '%.2f'), ...
    '&', num2str(temp2(1), '%.2f'), '&', num2str(temp2(2), '%.2f'), ...
    '&', num2str(temp2(3), '%.2f'), '&', num2str(temp2(4), '%.2f'), ...
    '&', num2str(temp2(5), '%.2f'), '&', num2str(temp2(6), '%.2f'), ...
    '&',num2str(temp2(7), '%.2f'), '\\*[-.60em]'])
disp(['&', num2str(mu_ui_ratio*adj_factor, '%.4f'), ...
    '&', num2str(temp3(1), '%.4f'), '&', num2str(temp3(2), '%.4f'), ...
    '&', num2str(temp3(3), '%.4f'), '&', num2str(temp3(4), '%.4f'), ...
    '&', num2str(temp3(5), '%.4f'), '&', num2str(temp3(6), '%.4f'), ...
    '&',num2str(temp3(7), '%.4f'), '\\*[-.60em]'])
```

```
save([output_folder,'/AARP_',sex_name,'_',data_name,...
    '_ui_food_trace.mat'],'-mat','usual_intake_food_trace');
save([output_folder,'/AARP_',sex_name,'_',data_name,...
    '_ui_energy_trace.mat'],'-mat','usual_intake_energy_trace');

save([output_folder,'/AARP_',sex_name,'_',data_name,...
    '_food_distribution.mat'],'-mat','food_distribution', 'food_p_mat');
save([output_folder,'/AARP_',sex_name,'_',data_name,...
    '_energy_distribution.mat'],'-mat','energy_distribution', ...
    'energy_p_mat');
save([output_folder,'/AARP_',sex_name,'_',data_name,...
    '_ratio_distribution.mat'],'-mat','ratio_distribution', 'ratio_p_mat');
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% What percentage of MCMC steps in which beta1 moves
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
beta1_accept_rate = beta1_accept_count/nMCMC;
disp(['beta1 accept rate is ', num2str(beta1_accept_rate*100), '%'])

toc
diary off

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Generate report
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
diary([output_folder,'/Output_MCMC_',dataset_name,'_', sex_name, '_', ...
    num2str(mmi), '_recalls_', with_covariates_method, '_',  ...
    datestr(now, 'mm_dd_yyyy'), '.txt']);
generate_latex_report(Sigmau_postmean, Sigmae_postmean, beta_postmean, ...
    Sigmau_postsd, Sigmae_postsd, beta_postsd, Corru_postmean, ...
    Corre_postmean, alpha_postmean, alpha_postsd, alpha_ci, ...
    never_postmean, never_postsd, never_ci, with_covariates_ind, ...
    with_covariates_method, dataset_name, data_name, sex_name, mmi, ...
    rw_ind, update_beta1_var_ind, lambda_ffq_food, lambda_ffq_energy, ...
    lambda_rec_food, lambda_rec_energy, beta1_accept_rate)
diary off
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Generate imaginary people's usual intake n_gen times
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
if n_gen > 0
    n_gen = ceil(n_gen);
    disp(['Start generating ', num2str(n_gen), ...
        ' realizations of usual intake'])
    data_wide = generate_usual_intake(Sigmau_postmean, Sigmae_postmean, ...
        beta_postmean, alpha_postmean, Xtildei, GGalpha, mumu, sigsig, ...
        mu_e, sig_e, a0_food, a0_energy, lambda_rec_food, ...
        lambda_rec_energy, n_gen);
    disp('Generating usual intake completed')
    UID = [1:n]';
    covariates = [UID, zz(:,[Age_ind, bmi_ind, ffq_food_ind, ...
        ffq_energy_ind])];
    data_long = NaN(n*n_gen, (size(covariates, 2)+2));
    for b = 1:n_gen
        data_long(b:n_gen:n*n_gen, : ) = ...
            [covariates, data_wide(:,[b, b+n_gen])];
    end
    data_wide = [covariates, data_wide];
    save([output_folder,'/AARP_',sex_name,'_',data_name, ...
        '_generated_usual_intake_', num2str(n_gen),'_realizations.mat'],...
        '-mat','data_wide', 'data_long');
%   data_wide:      generated data set with columns of the following order:
%                   ID, Age, bmi, ffq food, ffq_energy, recall food 1, ...,
```

```
%                   recall food n_gen, recall energy 1, ... ,
%                   recall energy n_gen
%   data_long:      same data set as data_wide, but stacked. The columns
%                   are ID, Age, bmi, ffq food, ffq_energy, the generated
%                   food and generated energy. The 1st row is recall 1
%                   for subject 1, the n_gen th row is recall n_gen
%                   for subject 1, the (n_gen+1)th row is the recall 1
%                   of subject 2, ...
disp(['The generated data set is saved as ', output_folder,'/AARP_', ...
   sex_name, '_',data_name,'_generated_usual_intake_', num2str(n_gen),...
   '_realizations.mat']);
disp('In that file, you will find two tables, data_wide and data_long.')
disp(['The columns of data_wide are ID, Age, bmi, ', data_name, ...
   ' from ffq, energy from ffq, ', data_name, ' from recall 1, ..., ', ...
   data_name, ' from recall ', num2str(n_gen), ...
   ', energy from recall 1, ... , energy from recall ',num2str(n_gen),'.'])
disp(['The columns of data_long are ID, Age, bmi, ', data_name, ...
   ' from ffq, energy from ffq, ', data_name, ...
   ' from recalls, energy from recalls.'])
end
```